# Machine Learning Methods for Classification Problems in Biomedical Signal Processing

Von der Fakultät Bau- und Umweltingenieurwissenschaften und
dem Stuttgart Center for Simulation Science
der Universität Stuttgart zur Erlangung der Würde
einer Doktor-Ingenieurin (Dr.-Ing.)
genehmigte Abhandlung

von

Emilie Ismailova

aus

Taschkent

Hauptberichter: Prof. Dr. Syn Schmitt
Mitberichterin: Prof. Lynne E. Bilston, PhD
Tag der mündlichen Prüfung: 05.11.2025

Institut für Modellierung und Simulation Biomechanischer Systeme
der Universität Stuttgart

2025

D 93
Institute for Modelling and Simulation of Biomechanical Systems
University of Stuttgart, Germany, 2025



ISBN 978-3-946412-98-4

# Declaration of Originality

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Stuttgart, November 10, 2025

Emilie Ismailova

*To Avas Vasilevich Khugaev, a brilliant physicist and mentor.*

# Acknowledgements

I would like to express my deep gratitude to everyone who has supported this work through their guidance, collaboration, and the sharing of data and expertise. The following contributors are listed in alphabetical order:

Prof. Lynne Bilston, PhD (University of New South Wales; Neuroscience Research Australia)
Diana Blazevska (Neuroscience Research Australia)
Jeroen Jeneson, PhD (University Medical Center Groningen)
Dr.-Ing. Thomas Klotz (Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart)
Fiona Knapman, PhD (Neuroscience Research Australia)
Zoia Lateva, PhD (VA Palo Alto Health Care System)
Kathy Lung (Neuroscience Research Australia)
PD Dr. med. Justus Marquetand (Eberhard Karls University of Tübingen; Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart)
Kevin C. McGill, PhD (Department of Functional Restoration, Stanford University)
Prof. Oliver Röhrle, PhD (Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart)
Prof. Dr.-Ing. Alina Roitberg (University of Hildesheim)
Prof. Dr. Syn Schmitt (Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart)
Dick Stegeman, PhD (Department of Neurology, Radboud University Medical Centre)

# Table of Contents

**Table of Contents** I

**List of Figures** VII

**List of Tables** X

**Deutsche Zusammenfassung** XI

**English Abstract** XV

**Acronyms** XIX

**Chapter I Motivation and Research Questions**

I.1 Skeletal Muscle Disorders Datasets ..... 2
I.1.1 Clinical Motivation ..... 2
I.1.2 State of the Art in Classification of Relevant Skeletal Muscle Disorders ..... 3
I.1.3 Limitations of Existing Approaches ..... 5
I.1.4 Research Questions ..... 6
I.2 Obstructive Sleep Apnea Dataset ..... 7
I.2.1 Clinical Motivation for Predicting the Outcomes of Treatment with Mandibular Advancement Splints ..... 7
I.2.2 State of the Art in Prediction the Outcomes of Treatment with Mandibular Advancement Splints ..... 7
I.2.3 Limitations of Existing Approaches ..... 9
I.2.4 Research Question ..... 10

**Chapter II Biomedical Signal Sources: Physiological Context and Clinical Measurement**

II.1 Skeletal Muscle Physiology and Relevant Pathophysiology ..... 11
II.1.1 Basic Architecture and Action Potential Generation ..... 12
II.1.2 Ion Channels and Muscle Fiber Electrophysiology ..... 12
II.1.3 Skeletal Muscle Disorders: Skeletal Muscle Channelopathies .. 13
II.2 Electromyography ..... 14
II.2.1 Fundamentals of Electromyographic Measurements ..... 15
II.2.2 Fibrillation Potentials ..... 16
II.2.3 Myotonic Discharges ..... 17
II.3 Obstructive Sleep Apnea ..... 19
II.3.1 Anatomy and Physiology of the Upper Airway Relevant for Development Obstructive Sleep Apnea ..... 21
II.3.2 Pathophysiology of Obstructive Sleep Apnea ..... 21
II.3.3 Polysomnography Overview and Its Role in Obstructive Sleep Apnea Diagnosis ..... 23
II.4 Mandibular Advancement Splints ..... 24
II.4.1 Treatment with Mandibular Advancement Splints in Obstructive Sleep Apnea ..... 25
II.4.2 Relevant Polysomnographic Signals and Their Significance for Prediction of Mandibular Advancement Splint Therapy Outcomes ..... 27

**Chapter III Data and Hypotheses**

III.1 Skeletal Muscle Disorders Datasets ..... 32
III.1.1 Skeletal Muscle Channelopathy Dataset ..... 32
III.1.2 Simulated Skeletal Muscle Channelopathy Dataset ..... 34
III.1.3 Skeletal Muscle Fibrillation Potentials Dataset ..... 36
III.1.4 Hypothesis for Skeletal Muscle Disorders Dataset ..... 36
III.2 Obstructive Sleep Apnea Dataset ..... 41
III.2.1 Obstructive Sleep Apnea Dataset ..... 41
III.2.2 Hypothesis for Obstructive Sleep Apnea Dataset ..... 43

**Chapter IV Signal Processing Methods and Analysis Techniques**

IV.1 Fundamentals of Biomedical Signal Processing ..... 48
IV.1.1 Discrete-Time Signals and Sampling Theorem ..... 48
IV.1.2 Linear Time-Invariant Systems ..... 49

IV.1.3 Time-Domain Characteristics of Signal ..... 50
IV.2 Noise Reduction and Digital Filtering ..... 53
IV.2.1 Signal-to-Noise Ratio ..... 53
IV.2.2 Digital Filters ..... 54
IV.2.3 Filtering Approaches in Digital Signal Processing ..... 55
IV.2.4 Advantages and Disadvantages of Filtering ..... 56
IV.3 Frequency-Domain and Spectral Analysis ..... 57
IV.3.1 Spectrum: Definition and Importance ..... 57
IV.3.2 Short-Time Fourier Transform and Spectrograms ..... 60
IV.3.3 Continuous Wavelet Transform ..... 63
IV.3.4 Choice of Mother Wavelet ..... 65
IV.3.5 Scalogram ..... 67
IV.4 Image Processing ..... 70
IV.4.1 Images as a Two-Dimensional Signals and Their Basic Properties 70
IV.4.2 Introduction to Texture Analysis ..... 71
IV.4.3 Grey-Level Co-Occurrence Matrix ..... 72
IV.5 Feature Extraction: Skeletal Muscle Disorders Datasets ..... 74
IV.5.1 Signal Quality and Processing ..... 75
IV.5.2 Time-domain Features ..... 81
IV.5.3 Features for Image-based Learning ..... 83
IV.5.4 Synthetic Skeletal Muscle Channelopathy Dataset ..... 85
IV.5.5 Skeletal Muscle Fibrillation Potentials Dataset ..... 85
IV.6 Feature Extraction: Obstructive Sleep Apnea Dataset ..... 86
IV.6.1 Signal Quality and Features for Non-image based learning ..... 87
IV.6.2 Features for Image-based Learning ..... 93
IV.6.3 Features for Non-image based learning obtained from Scalograms 95

**Chapter V Machine Learning Methods for Signal Classification**

V.1 Experimental Setup ..... 98
V.1.1 Objectives ..... 98
V.1.2 Cross-Validation, Data Splitting, and Model Assessment ..... 99
V.1.3 Evaluation Metrics ..... 101
V.2 Machine Learning Methods ..... 105
V.2.1 Classical Machine Learning Methods ..... 105
V.2.2 Neural Networks ..... 113
V.2.3 Transfer Learning and Neural Network Ensembles ..... 119
V.2.4 Data Augmentation Methods ..... 122
V.3 Skeletal Muscle Disorders Datasets: Classification ..... 125
V.3.1 Results of Classical Machine Learning Methods ..... 126

V.3.2 General Pipeline: Transfer Learning and Ensembles ..... 132
V.3.3 Results: Skeletal Muscle Channelopathies Dataset ..... 143
V.3.4 Initial Single-Model Test: Simulated Skeletal Muscle Channelopathy Dataset ..... 154
V.3.5 Results: Fibrillation Potentials Dataset ..... 158
V.4 Obstructive Sleep Apnea Dataset: Classification ..... 166
V.4.1 Application of Scalogram-Based Transfer Learning to Polysomnography Data Classification ..... 166
V.4.2 Statistical Analysis of Scalograms Based on Gray-Level Co-occurrence Matrix Features ..... 174
V.4.3 Application of Classical Machine Learning Methods for Classification of Obstructive Sleep Apnea Dataset ..... 177
V.4.4 Testing Random Forest Model on Downsampled Signals of Obstructive Sleep Apnea Dataset ..... 183

**Chapter VI Conclusion and Outlook**

VI.1 Conclusion: Skeletal Muscle Disorders Datasets ..... 188
VI.1.1 Discussion and Implications: Skeletal Muscle Channelopathy Dataset ..... 188
VI.1.2 Discussion of Fibrillation Potentials Dataset: Analysis of the Algorithm's Universal Applicability ..... 191
VI.1.3 Limitations and Future Research Directions ..... 192
VI.2 Conclusion: Obstructive Sleep Apnea Dataset ..... 193
VI.2.1 Discussion and Implications ..... 194
VI.2.2 Signal Quality and Its Importance for Classification ..... 196
VI.2.3 Limitations and Future Research Directions ..... 196

**References**

# List of Figures

## Chapter II Biomedical Signal Sources: Physiological Context and Clinical Measurement


Figure II.1 Pathophysiology of Myotonic Discharges ..... 20

Figure II.2 Inspiratory Airflow Morphologies ..... 29


## Chapter III Data and Hypotheses


Figure III.3 Modeled Myotonic Discharges for Chloride and Sodium Channel Defects ..... 36


## Chapter IV Signal Processing Methods and Analysis Techniques


Figure IV.4 Construction of a Scalogram of a Sample Signal ..... 69

Figure IV.5 The iEMG Signal and its Spectrogram ..... 75

Figure IV.6 Close-up View of the Second Discharge of the iEMG Signal .. 76

Figure IV.7 Power Spectral Density Comparison of Discharge and Noise .. 78

Figure IV.8 `p54CR_L_BICEP_2.csv` iEMG Signal and its Spectrogram ..... 79

Figure IV.9 Close-up View of a Train of Events in `p54CR_L_BICEP_2.csv` . 80

FIGURE IV.10 Comparison of Averaged Power Spectral Densities for Sodium and Chloride Channelopathy Classes ..... 80
FIGURE IV.11 Spectrogram and Scalogram Comparison of `p59SD_R_RECT_FEM_1.csv` ..... 83
FIGURE IV.12 Time-Amplitude and Time-Frequency Domain Representations of a Voluntary Action Potential Trains iEMG Recording ..... 86
FIGURE IV.13 Time-Amplitude and Time-Frequency Domain Representations of a Fibrillation Potential iEMG Recording ..... 87
FIGURE IV.14 Comparison of Power Spectral Densities for Patient Airflow in OSAMAS Dataset ..... 90
FIGURE IV.15 Comparison of Power Spectral Densities for Patient Airflow in CRC Dataset ..... 91
FIGURE IV.16 Comparison of Power Spectral Densities for Patient Airflow in PhysMAS Dataset ..... 92
FIGURE IV.17 Comparison of Hypopnea Events between OSAMAS_012 and OSAMAS_029 ..... 94

## CHAPTER V MACHINE LEARNING METHODS FOR SIGNAL CLASSIFICATION

FIGURE V.18 Correlation Heatmap of EMG Features ..... 127
FIGURE V.19 Precision-Recall Curves of Four Different Classical Machine Learning Models ..... 128
FIGURE V.20 Receiver Operating Characteristic Curves of Four Different Classical Machine Learning Models ..... 129
FIGURE V.21 Confusion Matrices of Four Different Classical Machine Learning Models ..... 130
FIGURE V.22 Aggregated Classification Performance of the InceptionResNetV2 on Morse Wavelet Transformed Dataset Averaged across Splits ..... 146
FIGURE V.23 Test-set Performance for InceprtionResNetV2 on Morse Wavelet Transformed Dataset on Splits 3 and 5 ..... 146
FIGURE V.24 Grad-CAM for Correctly Classified Chloride Myotonias ..... 148
FIGURE V.25 Grad-CAM for Correctly Classified Sodium Myotonias ..... 148
FIGURE V.26 Discriminative Micro-patterns Revealed by Grad-CAM ..... 149
FIGURE V.27 Grad-CAM for Three Misclassified Recordings ..... 150

Figure V.28 Aggregated Ensemble Performance and Uncertainty ..... 152
Figure V.29 Ensemble Performance and Uncertainty on Split 3 ..... 153
Figure V.30 Ensemble Performance and Uncertainty on Split 5 ..... 153
Figure V.31 Correct and Incorrect Classifications per Split ..... 154
Figure V.32 Grad-CAM Saliency Maps for Healthy Controls ..... 156
Figure V.33 Grad-CAM Saliency Maps for Chloride-channel Myotonia .... 156
Figure V.34 GradCAM Saliency Maps for Sodium-channel Myotonia ..... 157
Figure V.35 Representative Grad-CAM Overlays for Patient 34 ..... 157
Figure V.36 Aggregated Performance of the InceptionResNetV2 on the Morse Wavelet Transformed Dataset ..... 160
Figure V.37 Grad-CAM for Fibrillation and Voluntary Action Potentials . 161
Figure V.38 Grad-CAM for Misclassified Samples in Fibrillations Dataset . 161
Figure V.39 Aggregated Ensemble Performance and Uncertainty on Fibriallation Dataset ..... 164
Figure V.40 Ensemble Performance and Uncertainty on Split 4 of Fibrillations Dataset ..... 165
Figure V.41 Ensemble Performance and Uncertainty on Split 2 of Fibrillations Dataset ..... 165
Figure V.42 Adaptive Pooling Pipeline Results ..... 169
Figure V.43 Training and Validation Loss across Folds for Baseline Model 170
Figure V.44 Training and Validation Loss across Folds under the Feature Fusion Pipeline ..... 171
Figure V.45 GLCM Feature Distributions for OSAMAS Dataset ..... 175
Figure V.46 GLCM Feature Distributions for CRC Dataset ..... 176
Figure V.47 GLCM Feature Distributions for PhysMAS Dataset ..... 176
Figure V.48 Random Forest Classifier: Full, Time Domain and Spectral Features Comparison ..... 180
Figure V.49 Modified Random Forest Classifier Results on Combined CRC and PhysMAS Dataset ..... 182
Figure V.50 Feature Importance Plot for Random Forest Model ..... 183
Figure V.51 Random Forest Model Metrics on Downsampled Data ..... 185
Figure V.52 Random Forest Metrics on Combined Downsampled Data .... 186

# List of Tables

## Chapter III Data and Hypotheses


Table III.1 Baseline Parameters for the Channelopathy Simulation ..... 35

Table III.2 Synchronized Flow Data Overview ..... 43

Table III.3 Overview of Flow Data Files ..... 43


## Chapter V Machine Learning Methods for Signal Classification


Table V.1 Metrics for Classical Machine Learning Models on Full Feature Set ..... 130

Table V.2 Metrics for Classical Machine Learning Models on Reduced Feature Set ..... 131

Table V.3 Classification Metrics for Loss Functions on Test Set ..... 135

Table V.4 Validation and Test Metrics for Data Augmentation Methods in iEMG Signal Classification Averaged over Five Splits ..... 138

Table V.5 Averaged Metrics over Five Splits for Individual Classifiers ... 145

Table V.6 Metrics Summary for InceptionResNetV2 (Morse) on Splits 3 and 5 ..... 145

Table V.7 Aggregated Metrics for Ensemble Predictions across Splits ... 151

Table V.8 Metrics for Ensemble Predictions for Selected Splits ..... 151

Table V.9 Aggregated Validation and Test Metrics ..... 155

Table V.10 Averaged Metrics over Splits for InceptionResNetV2 and MaxViT Classification of Fibrillation Potentials ..... 159
Table V.11 Aggregated Metrics for Ensemble Classification of Fibrillation Potentials across Splits ..... 163
Table V.12 Metrics for Ensemble Predictions for Splits 2 and 4 ..... 164
Table V.13 Classification Metrics for Baseline, Adaptive Pooling, Feature Fusion, and Fine-Tuning, Simple CNN with Feature Fusion Pipelines on Validation Set ..... 168
Table V.14 Average Classification Metrics for Random Forest on Individual Datasets, CRC and PhysMAS Combined, and All Three Datasets Combined ..... 178
Table V.15 Average Classification Metrics for Random Forest and Random Forest Ensembled with SVM on CRC and PhysMAS Combined Dataset ..... 181
Table V.16 Average Classification Metrics for Random Forest with downsampled CRC and PhysMAS and OSAMAS Datasets ... 185
Table V.17 Average Classification Metrics for Random Forest with Downsampled CRC and PhysMAS and OSAMAS Datasets .. 186

# Deutsche Zusammenfassung

Routinephysiologische Daten enthalten wertvolle diagnostische Hinweise, dennoch verlassen sich Kliniker größtenteils auf die manuelle visuelle Inspektion roher Signalverläufe. Dieser subjektive Ansatz berücksichtigt nicht die informativen Muster, die im Zeit-Frequenz-Bereich der Signale verborgen sind. In dieser Arbeit untersuchen wir, ob moderne maschinelle Lernverfahren, angewendet auf spektrale Darstellungen biomedizinischer Signale, latente Biomarker identifizieren und in klinisch verwertbare Erkenntnisse umwandeln können. Das Projekt adressiert diese Fragestellung anhand zweier unterschiedlicher Signaltypen: (i) elektrische Signale, indem intramuskuläre Elektromyographie (iEMG) zur Klassifikation spontaner Skelettmuskelaktivität genutzt wird, und (ii) mechanische Signale, durch die Vorhersage des Therapieerfolgs bei obstruktiver Schlafapnoe mittels nasaler Luftstrommessungen, die während der Polysomnographie erhoben werden.

Der erste Teil der Arbeit untersucht Skelettmuskel-Kanalopathien – eine Gruppe neuromuskulärer Erkrankungen, welche die Erregbarkeit der Zellmembran stören und klinisch als Myotonie sichtbar werden. Die genetische Ursache einiger dieser Erkrankungen lässt sich auf Mutationen in den Genen SCN4A oder CLCN1 zurückführen, welche jeweils Natrium, bzw. Chloridkanäle kodieren. Diese Kanäle sind spezialisierte Membranproteine, die bei der Entstehung und Weiterleitung von Aktionspotentialen eine entscheidende Rolle spielen. Defekte dieser Kanäle führen typischerweise zu einer pathophysiologischen Übererregbarkeit der Muskeln, die auf iEMG-Aufzeichnungen als myotone Entladungen sichtbar wird. Es herrscht jedoch Uneinigkeit darüber, ob sich aus den Eigenschaften der myotonen Entladungen der zugrundeliegende Kanaldefekt, beispielsweise Natrium- versus Chloridkanaldefekt, differenzieren lässt. Dies führt zur weiterführenden Frage, ob eine stabile Beziehung zwischen Genotyp und EMG-Phänotyp bei Muskelkanalopathien existiert.

Die präzise Identifikation des zugrundeliegenden Ionenkanaldefekts ist für eine maßgeschneiderte Therapie und prognostische Einschätzung entscheidend. Aktuell

verlassen sich Kliniker ausschließlich auf genetische Tests zur Differenzierung zwischen Natrium- und Chloridkanaldefekten. Obwohl detaillierte manuelle iEMG-Analysen defektspezifische Muster aufdecken können, ist dieser Prozess für den klinischen Alltag zu zeitaufwendig und komplex.

Unsere Studie zeigte die Existenz charakteristischer spektraler Merkmale in myotonen Entladungen bei Patienten mit Natrium- und Chloridkanaldefekten und adressierte den Bedarf an automatisierter Klassifikation dieser Merkmale. Hierzu transformierten wir iEMG-Aufzeichnungen mittels Wavelet-Transformation in spektrale Darstellungen (Skalogramme), welche anschließend durch ein Ensemble vortrainierter tiefer neuronaler Netzwerke klassifiziert wurden. Das resultierende Ensemble erzielte auf unbekannten Testdaten eine balancierte Genauigkeit von etwa 81 % und einen Brier-Score von 0,14. Eine selektive Vorhersageanalyse zeigte weiterhin, dass bei hohen Konfidenzschwellenwerten (größer als 0,85) die Modellgenauigkeit 90 % überschritt. Diese Ergebnisse verdeutlichen die potenzielle klinische Nützlichkeit dieser Methode zur Verbesserung der diagnostischen Effizienz, insbesondere durch Priorisierung genetischer Tests auf spezifische Mutationen.

Ein Ziel war es, spezifische physiologische Signalcharakteristiken für jeden Subtyp der Ionenkanaldefekte zu identifizieren. Elektrophysiologische Analysen offenbarten, wie unterschiedliche Kanaldefekte charakteristische Entladungsmuster erzeugen und so die zugrundeliegenden Mechanismen der Myotonie verdeutlichen. Gradientengewichtete Salienzanalysen identifizierten diskriminative spektrale Merkmale, darunter breitbandige, frühe Energiespitzen bei Chloridkanaldefekten sowie nachhaltige hochfrequente spektrale Komponenten bei Natriumkanaldefekten.

Um die physiologische Interpretation unserer Ergebnisse zu unterstützen, generierten wir synthetische myotone Entladungen mithilfe eines biophysikalischen Modells von Klotz et al. (2020). Das konvolutionale neuronale Netzwerk erkannte klassenspezifische spektrale Unterschiede in diesen simulierten Signalen. Dies stärkt die physiologische Relevanz der spektralen Motive, die als Marker für spezifische Kanalopathien identifiziert wurden, und erhöht somit die Interpretierbarkeit unseres Deep-Learning-Modells.

Zur weiteren Überprüfung der Robustheit und allgemeinen Anwendbarkeit des Ansatzes nutzten wir dieselbe Klassifikationspipeline, um willkürliche motorische Einheitaktivität von spontanen Fibrillationspotentialen zu unterscheiden – einem elektrophysiologischen Kennzeichen degenerativer und neurogener Muskelerkrankungen. Ohne Änderung der Netzwerkarchitektur oder Hyperparameter erzielte das Ensemble weiterhin eine hohe Leistung, mit einer balancierten Genauigkeit von 87 % und einem Brier-Score von 0,11. Diese Ergebnisse bestätigen die Übertragbarkeit des Ansatzes über verschiedene Pathologien hinweg und belegen dessen Vielseitigkeit für die automatisierte EMG-Diagnostik.

Der zweite Teil der Studie untersucht die obstruktive Schlafapnoe, eine häufige chronische Erkrankung, die oft mit mandibulären Protrusionsschienen behandelt wird. Ungefähr ein Drittel der Patienten spricht jedoch nicht ausreichend auf die Therapie an. Da die Vorhersage des Therapieerfolgs basierend auf der Baseline-Polysomnographie weiterhin schwierig ist, untersuchten wir, ob das routinemäßig aufgezeichnete nasale Luftstromsignal bereits Merkmale enthält, die den Therapieerfolg vorhersagen können.

Die spektralen Merkmale des Signals wurden quantifiziert, indem dominante Frequenzen im niedrigen, mittleren und hohen Frequenzbereich extrahiert wurden. Klassische Machine-Learning-Algorithmen, insbesondere Random-Forest-Klassifikatoren, zeigten bei Nutzung hochfrequent abgetasteter Signale vielversprechende Ergebnisse. Diese Klassifikatoren identifizierten Responder mit einer Recall-Rate von fast 90 % und einem Cohen's Kappa von ca. 0,48. Eine Reduktion der Abtastfrequenz beeinträchtigte hingegen deutlich die Vorhersageleistung, was die zentrale Rolle mittlerer und hoher Frequenzen bei dieser Klassifikationsaufgabe unterstreicht.

Diese Studie stellt erstmalig eine automatisierte, unsicherheitsbewusste Diagnostikpipeline vor, die Muskelkanalopathien direkt aus Routine-EMG-Signalen unterscheidet und durch spektrale Signalanalyse und maschinelles Lernen breite Anwendbarkeit in verschiedenen biomedizinischen Kontexten verspricht.

# English Abstract

Routine physiological data contains rich diagnostic cues, yet clinicians still rely chiefly on manual visual inspection of raw waveforms. This subjective approach does not consider informative patterns hidden in the time-frequency domain of the signal. Here, we ask whether modern machine-learning algorithms, applied to spectral representations of biomedical signals, can uncover latent biomarkers and turn them into actionable clinical insights. The project addresses this overarching question by focusing on two distinct signal types: (i) electrical by classifying intramuscular electromyography (iEMG) to distinguish spontaneous skeletal muscle activity, and (ii) mechanical by predicting treatment outcomes in obstructive sleep apnea from nasal airflow recordings obtained during polysomnography.

First part of this work investigates skeletal muscle channelopathies – a group of neuromuscular disorders that disturb the cell membrane excitability, which results clinically in myotonia. The genetic aetiology of some of these disorders can be traced to mutations in the SCN4A or CLCN1 genes, which encode sodium and chloride channels, respectively. These channels are specialized proteins in the cell membrane that play a crucial role in generation and propagation of action potentials. Commonly, sodium or chloride channel defects lead to pathophysiological hyperexcitability of muscles, which is observable as myotonic discharges on iEMG recordings. However, there is an ongoing debate about whether the properties of the myotonic discharge can differentiate the type of channel defect, such as sodium versus chloride. This discussion leads to the broader question of whether a stable genotype-to-EMG-phenotype relationship exists in muscle channelopathies.

Accurate identification of the underlying ion-channel defect is also essential for tailored treatment and informed prognosis. At present, clinicians depend exclusively on genetic testing to distinguish sodium- from chloride-channel defects. Although detailed, manual inspection of iEMG recordings can potentially reveal defect-specific patterns in research settings, the procedure is too time-consuming and complex for routine clinical practice.

Our study demonstrated the existence of distinct spectral features in myotonic discharges of patients with sodium and chloride channel defects and addressed the need for their automated classification. We developed and validated a method for this purpose, transforming iEMG recordings into their spectral representations (scalograms) via wavelet transform. These scalograms were subsequently classified using an ensemble of pre-trained deep neural networks. The resulting ensemble achieved a balanced accuracy of approximately 81 % and a Brier score of 0.14 on unseen test data. A selective-prediction analysis further indicated that at high-confidence thresholds (greater than 0.85), the model's accuracy exceeded 90 %. These results show the potential clinical utility of this approach for enhancing diagnostic efficiency, specifically by helping to prioritize genetic testing for a specific mutation.

One of our objectives was to identify physiological signal characteristics specific to each subtype of ion-channel defect. Electrophysiological analysis can reveal how distinct channel defects produce characteristic discharge patterns, clarifying the underlying mechanisms of myotonia. Gradient-weighted saliency mapping identified discriminative spectral features, including broadband, early-burst energy characteristic of chloride-channel defects, while sustained high-frequency spectral components were observed in sodium-channel defect class samples.

To support the physiological interpretation of our findings, we generated synthetic myotonic discharges using the biophysical model developed by Klotz et al. (2020). The convolutional neural network detected class-specific spectral differences within these simulated signals. This reinforces the physiological relevance of the spectral motifs identified as markers for distinct channelopathies, thereby enhancing the interpretability of our deep learning model.

To further test the robustness and general applicability of the framework, we used the same classification pipeline to distinguish voluntary motor unit activity from spontaneous fibrillation potentials – an electrophysiological hallmark of degenerative and neurogenic muscle disease. Without altering any network architecture or hyperparameters, the ensemble maintained high performance, achieving a balanced accuracy of 87 % and a Brier score of 0.11. These findings confirm that the spectral feature learning approach transfers across various pathologies and could therefore serve as a versatile tool for automated EMG diagnostics.

The second part of this study investigates obstructive sleep apnea, a common chronic disorder often treated with mandibular advancement splints. Roughly one-third of patients, however, do not respond adequately to this treatment. Because it is still difficult to predict patient responsiveness from baseline polysomnography, we examined whether the routinely recorded baseline nasal-airflow signal contains features that can predict treatment outcome.

The signal's spectral characteristics were quantified by extracting dominant frequencies within low, mid, and high-frequency bands. Using these features, classical machine learning algorithms, particularly Random Forest classifiers, demonstrated promising results. Specifically, when trained on signals sampled at sufficiently high frequencies, these classifiers accurately identified responders with a recall approaching 90 % and a Cohen's kappa of approximately 0.48. Conversely, downsampling the data to lower frequencies significantly impaired predictive performance, highlighting the critical role of mid- and high-frequency spectral content in this classification task.

This study introduces the first automated, uncertainty-aware diagnostic pipeline capable of distinguishing skeletal muscle channelopathies directly from routine EMG signals. By employing spectral signal characteristics within a machine learning framework, it suggests broad applicability across various biomedical contexts. We developed a purely signal-based predictive tool with good clinical diagnostic potential.

# Acronyms

| Acronym | Description |
|---|---|
| 2D | two-dimensional |
| 3D | three-dimensional |
| ACh | acetylcholine |
| AHI | apnea-hypopnea index |
| CNN | convolutional neural network |
| CPAP | continuous positive airway pressure |
| DFT | discrete Fourier transform |
| DSP | digital siganl processing |
| DTFT | discrete-time Fourier transform |
| ECG | electrocardiography |
| EEG | electroencephalography |
| EMG | electromyography |
| EVL | enhanced waveform length |
| iEMG | intramuscular electromyography |
| sEMG | surface electromyography |
| EOG | electrooculography |
| LCOV | log coefficient of variation |
| LTI | linear time-invariant |
| MAS | mandibular advancement splints |
| MAV | mean absolute value |
| ME | average energy |

| | |
|---|---|
| MFL | maximum fractal length |
| MU | motor unit |
| MUAP | motor unit action potential |
| MUAPs | motor unit action potentials |
| NED | negative effort dependence |
| NN | neural networks |
| non-REM | non-rapid eye movement |
| OSA | obstructive sleep apnea |
| PSD | power spectral density |
| PSG | polysomnography |
| REM | rapid eye movement |
| RMS | root-mean-square |
| SR | sarcoplasmic reticulum |
| SSC | slope sign changes |
| std | standard deviation |
| STFT | short-time Fourier transform |
| SVM | support vector machines |
| TM | temporal moment |
| VAR | variance |
| ZC | zero crossings |
| XGBoost | eXtreme gradient boosting |

## Chapter I

# Motivation and Research Questions

Continuous physiological monitoring generates large amounts of biomedical waveforms, yet clinicians typically review them only in the raw time-amplitude domain. This manual inspection is limited by human perception and prone to inter-observer variability. Spectral representations of the biomedical signals might offer a more objective perspective. Decomposing a signal into its frequency components can reveal patterns and energy distributions that remain hidden in the time series. We therefore argue that spectra of routine signals may harbor latent biomarkers.

This study tests the hypothesis that subtle spectral signatures in biomedical signals carry clinically relevant information in two distinct contexts. We first analyze needle electromyography (EMG) recordings – direct measurements of skeletal-muscle electrical activity, to determine whether changes in their spectral evolution reflect underlying pathophysiology. We then examine nasal-airflow signals collected during overnight polysomnography, using spectral features observed during flow-limited breaths to predict the efficacy of a targeted therapeutic device for obstructive sleep apnea. In both settings, the overarching goal is to connect spectral signal characteristics with physiological mechanisms and clinical outcomes.

We assume that machine learning models trained on spectral features can learn representations that transfer across signal modalities: electrical (EMG) and mechanical (nasal airflow). Accordingly, we pursue three objectives: (i) automate electromyographic phenotype classification, (ii) predict treatment efficacy for obstructive sleep apnea, and (iii) evaluate the cross-modal applicability of the spectral and machine learning methods. Achieving these aims would not only streamline two diagnostic pathways but also demonstrate how

spectral features can uncover latent information in heterogeneous clinical waveforms.

## I.1 Skeletal Muscle Disorders Datasets

### I.1.1 Clinical Motivation

Non-dystrophic myotonias are an important group of skeletal muscle channelopathies characterized by delayed muscle relaxation – myotonia, due to altered membrane excitability (Morales et al., 2020; Cannon, 2015). While a variety of ion channels can be affected in different channelopathies, non-dystrophic myotonias are primarily caused by mutations in two genes: CLCN1, encoding the skeletal muscle chloride channel, and SCN4A, encoding the voltage-gated sodium channel. Mutations in these genes make muscle fibers hyperexcitable, leading to repetitive firing and clinical myotonia (Morales et al., 2020; Cannon, 2015).

Patients with chloride-channel myotonia and sodium-channel myotonia exhibit a broad and overlapping spectrum of phenotypes, varying in age at onset, distribution and severity of stiffness, presence of muscle hypertrophy, and episodes of weakness (Drost et al., 2015; Vereb et al., 2021). This overlap often makes it difficult to distinguish the underlying molecular defect on the basis of clinical examination alone. Indeed, genetic testing has revealed that roughly 20 % of individuals clinically presumed to have CLCN1-related myotonia actually harbor SCN4A mutations (Trip et al., 2008). Such misclassification has direct consequences for both therapy and prognosis.

Accurate diagnosis of whether a patient's myotonic symptoms arise from a chloride or sodium channel mutation has important implications for therapy and prognosis (Morales et al., 2020). While both subtypes share the symptom of myotonic stiffness, there are differences in triggers and potential complications. For example, sodium-channel related myotonias are highly sensitive to environmental factors like cold exposure and exercise or elevated serum potassium levels and frequently include transient weakness or paralysis in addition to stiffness, whereas chloride-channel myotonias typically lack episodic paralysis and are less influenced by external conditions (Cannon, 2015).

Knowing the genetic subtype therefore guides patient advice: for instance, SCN4A patients are counseled to avoid cold and extreme hyperkalemic triggers, whereas such measures are less relevant in CLCN1 myotonia. Prognostic expectations also differ: certain SCN4A mutations can cause more severe phenotypes (even neonatal life-threatening myotonia in rare cases) (Matthews et al., 2010), whereas the spectrum of CLCN1 mutations ranges from mild stiffness to more disabling recessive forms, but generally without progressive weakness.

From a treatment perspective, both chloride and sodium channel myotonias often respond to sodium channel blockers like mexiletine which reduce membrane excitability (Matthews

et al., 2010). However, some sodium-channel myotonia patients with episodic weakness may benefit from acetazolamide (Markhorst et al., 2014) or lifestyle adjustments that would not apply to chloride-channel myotonia (Matthews et al., 2010). Thus, establishing the correct molecular diagnosis is important not only for symptomatic treatment but also for counseling on triggers, inheritance patterns, and long-term outlook (Morales et al., 2020; Trip et al., 2008).

In settings with limited resources, identifying the a chloride-channel defect by means of needle EMG can justify fast, single-gene testing for CLCN1, thereby cutting both diagnostic time and expense. Moreover, when EMG findings align with genetic test results, especially for variants of uncertain significance, they strengthen the clinical interpretation of the molecular data. Although genetic sequencing remains the definitive diagnostic standard, incorporating an EMG-based classifier could provide a rapid, cost-effective means to enhance diagnostic accuracy and inform therapeutic decisions. The primary clinical motivation of this study is therefore to develop an automated classifier that minimizes the need for specialized expertise in signal processing and seamlessly supports clinicians in their decision-making.

Building on this framework, we extend our machine-learning approach to the automated detection of spontaneous fibrillation potentials – an electrophysiological sign of neurogenic and degenerative muscle disorders. Manual identification of fibrillations relies on expert visual inspection of needle EMG signals, which can become especially challenging when fibrillation potentials occur concurrently with voluntary motor unit (MU) activity. In such cases, experts must discern irregular, single-fiber discharges superimposed on a train of voluntary potentials: a process that is laborious, subjective, and prone to inter-rater variability (Nandedkar et al., 2013). By standardizing detection of fibrillation potentials through automated algorithms, we aim to reduce observer bias, accelerate interpretation, and alleviate clinician workload, thus enhancing both the consistency and efficiency of EMG-based diagnosis.

### I.1.2 State of the Art in Classification of Relevant Skeletal Muscle Disorders

The diagnosis of non-dystrophic myotonias traditionally relies on clinical history, genetic testing, and characteristic EMG findings. On resting needle EMG, patients exhibit characteristic myotonic discharges: runs of spontaneous muscle-fiber action potentials that reflect membrane hyperexcitability due to defects in the skeletal muscle chloride or sodium channels.

Fournier et al. (2004) first demonstrated that tailored EMG protocols could distinguish these channelopathy subtypes. In their study of 51 genetically confirmed patients, they applied standardized short- and long-exercise tests to monitor compound muscle

action potential (CMAP) changes, defining five response patterns that discriminated CLCN1-related myotonia congenita from SCN4A-related myotonias. A subsequent protocol combining repeated short exercise with muscle cooling further enhanced genotype classification by identifying paramyotonia congenita in cases with normal room-temperature responses (Fournier et al., 2006). More recently, (Fournier et al., 2019) reviewed the clinical electrophysiology of muscle channelopathies and reiterated that specific patterns measured by needle EMG and muscles involved can provide etiological clues. In practice, however, these approaches still rely on expert interpretation of EMG waveforms under controlled conditions.

Drost et al. (2015) conducted blinded needle EMG in 66 genetically confirmed patients, recording myotonic discharges from five muscles per patient to extract individual myotonic trains and compute firing-pattern metrics. Remarkably, the first inter-discharge interval alone, consistently exceeding 30 ms in chloride-channel cases and falling below 30 ms in sodium-channel cases, provided over 95 % accuracy in genotype classification using a 30 ms threshold (Drost et al., 2015). Physiologically, this finding implies that in chloride-channel channelopathy, myotonic discharge begins at a lower firing rate with more gradual waxing and waning of burst frequency, whereas in sodium-channel myotonia, discharge initiates at a higher rate. Although such manual feature extraction can yield high diagnostic performance, it is labor-intensive and limited to a few predefined measures, potentially overlooking high-dimensional patterns that automated analysis might capture.

Initial biophysical modeling by Cannon et al. (1993) using a two-compartment skeletal-muscle membrane showed that impaired fast inactivation of voltage-gated sodium channels is sufficient to generate self-sustained trains of action potentials. Subsequent experimental and modeling studies by Cannon (1996b) and Vedantham et al. (2000) demonstrated that the magnitude of the persistent sodium current and the kinetics of slow inactivation together determine action-potential duration and repetitive firing patterns, thereby directly linking specific waveform features to underlying channel-gating defects.

The signal's spectrum offers valuable insights into its characteristics. One effective way to exploit these spectral features is through analysis of various spectral representations. In related domains such as ECG and EEG, researchers have leveraged pre-trained convolutional neural networks on signal time-frequency domain images to classify signal patterns (Salem et al., 2018; Benmalek et al., 2022; Pattnaik et al., 2024; Sun et al., 2022; Sun et al., 2025). This transfer-learning approach enables the use of powerful deep networks even with limited datasets. However, to our knowledge, no published study has applied scalogram-based deep learning specifically to needle EMG data; the only related work to date used surface EMG for gait analysis (Negi et al., 2023).

Researchers have applied various signal processing techniques to extract time- and

frequency-domain features from needle EMG waveforms, parameters such as potential duration, amplitude, rise time, phase count, and firing-interval regularity, to distinguish pathological fibrillation potentials from voluntary motor-unit potentials. Fibrillation potentials tend to have shorter durations and lower amplitudes than motor-unit potentials, and may show an initial positive deflection, uncommon in normal voluntary trains (Davalos et al., 2018). Nam et al. (2019) converted EMG waveforms into images and trained an Inception-v4 to classify positive sharp waves, fibrillations, and other spontaneous discharges, reaching about 93.8 % accuracy. More recently, Mel-spectrogram conversion combined with data augmentation and transfer learning, has further improved classification of resting EMG discharges (Nodera et al., 2019).

### I.1.3 Limitations of Existing Approaches

While prior studies demonstrate that EMG signals contain relevant biomarkers for differentiating myotonic disorders, the clinical translation of these findings remains hampered by methodological and practical limitations. Traditional EMG analysis techniques, including the specialized exercise tests of Fournier et al. (2004) and the discharge interval measurements Drost et al. (2015) require considerable expert involvement and carefully controlled conditions. For example, the near-perfect discrimination achieved by Drost et al. (2015) was specific to analyzing discharges from the rectus femoris muscle under standardized settings, with manual signal decomposition. Such rigorously controlled protocols are often impractical in routine clinical settings, where optimal muscle selection and clean recordings may not be feasible.

Furthermore, the analysis of Drost et al. (2015) relied heavily on manual identification of myotonic trains and excluded complex or short bursts, potentially biasing the results toward ideal signal morphology. Thus, methods that rely on a single muscle or a single hand-crafted feature might not generalize across patients, centers, or EMG systems. As person-to-person variability, due to factors such as age, disease severity, or fiber excitability, can possibly alter discharge characteristics, morphological overlap in EMG waveforms can lead to misclassification if only simple metrics are used.

Moreover, the pathophysiological assumptions underlying the criterion based solely on the initial interdischarge interval may oversimplify the complex membrane dynamics involved. The assumption that the initial discharge interval directly reflects resting membrane properties has not been experimentally verified, and the substantial variability observed in discharge amplitudes and waveform shapes remains poorly understood. Giant myotonic discharges exceeding 10 mV have been attributed to ephaptic coupling of neighbouring fibers, but this mechanism remains speculative and untested in controlled experiments (Drost et al., 2015).

Moreover, reliance on manually engineered features can introduce bias and may limit the

discovery of novel, discriminative signal attributes. Features effective in one dataset may under-perform when applied to data from other acquisition systems or clinical contexts. There is a clear need for approaches that can automatically learn robust representations from raw or minimally processed EMG signals: a space where deep learning holds substantial promise. However, deep models introduce their own challenges, including the need for interpretability and reliable generalization to unseen patients. To date, no fully automated tool is in clinical use for EMG-based genotype classification in myotonic disorders; diagnosis still depends primarily on genetic testing following clinical suspicion (Trip et al., 2008). This is an important translational gap that improved algorithmic solutions could help bridge, offering earlier and more accessible support for subtype differentiation in muscle channelopathies.

While existing EMG-based approaches clearly demonstrate the diagnostic potential for distinguishing between myotonia subtypes, their practical application remains hindered by the heavy dependence on expert-defined procedures and limited generalizability. Addressing these limitations through robust and fully automated classification algorithms, capable of processing heterogeneous needle EMG recordings with minimal expert involvement, remains a key research objective. Our goal is to develop such an algorithm that not only accurately classifies myotonic discharges but can also be generalized to other spontaneous muscle activities such as fibrillation potentials, thus advancing clinical diagnostic capabilities.

### I.1.4 Research Questions

This study aims to develop a machine learning model capable of distinguishing between chloride-channel and sodium-channel defects in patients diagnosed with non-dystrophic myotonic disorders. The focus is on capturing and learning the hypothetical differences in the morphology and dynamics of myotonic discharges, thereby providing a practical tool to aid in the differential diagnosis of myotonias.

Additionally, the study seeks to evaluate the model's ability to generalize beyond myotonic discharges to detect other pathological EMG activity, demonstrating broader applicability. Specifically, the study will assess whether the model can distinguish voluntary muscle contractions from fibrillation potentials, representing a different EMG abnormality.

By challenging the model to recognize fibrillation potentials, the research probes the adaptability of the approach to diverse clinical situations. This serves as a stress test for the universality of the method, examining whether a model developed for one category of abnormal EMG activity (myotonic discharges) can transfer its knowledge to another category (fibrillations) with minimal adjustments.

The anticipated contributions of this research are both clinical and technical. Clinically, the study aims to enhance differentiation of muscle channelopathies from routine med-

ical recording, potentially leading to earlier and more accurate diagnoses. Technically, the research applies state-of-the-art machine learning techniques to the classification of intramuscular EMG signals, bringing in new methods into the biomedical signal analysis.

## I.2 Obstructive Sleep Apnea Dataset

Obstructive sleep apnea (OSA) is a common chronic disorder characterized by repeated upper airway obstructions during sleep, that often requires long-term management. Mandibular advancement splints (MAS), a type of oral appliance therapy, have emerged as a primary alternative to continuous positive airway pressure (CPAP) for patients with OSA, particularly for those who are unable to tolerate CPAP (Chan et al., 2020; Sutherland et al., 2011). These devices mechanically advance the mandible during sleep, thereby enlarging and stabilizing the upper airway. Currently, MAS is recommended as a first-line treatment for mild-to-moderate OSA and is even considered in some severe cases when CPAP is not a viable option.

### I.2.1 Clinical Motivation for Predicting the Outcomes of Treatment with Mandibular Advancement Splints

A major challenge of MAS therapy is its highly variable efficacy among patients. While some individuals experience significant improvements, others exhibit minimal or no change in disease severity. In fact, over one-third of OSA patients show little or no reduction in severity following oral appliance therapy (Chen et al., 2020). Furthermore, a survey revealed that 88% of sleep physicians consider the unpredictable performance of oral appliance therapy a major obstacle to its widespread use (McCormack, 2022). Accurately predicting MAS treatment outcomes in advance would allow clinicians to tailor therapy: patients likely to benefit from MAS could be promptly directed to this treatment, while those unlikely to respond could be offered alternative options. This stratification would enhance treatment efficiency, improve patient satisfaction, and increase clinician confidence.

### I.2.2 State of the Art in Prediction the Outcomes of Treatment with Mandibular Advancement Splints

To address the clinical need for better patient selection, researchers have explored various predictors of MAS treatment outcomes. Traditional studies have focused on patient demographics, anthropometric measurements, and baseline OSA severity. For example, younger age, female gender, lower body mass index (BMI), smaller neck circumference, and less severe baseline apnea-hypopnea index (AHI) have been linked to more favorable treatment outcomes in several studies (Sutherland et al., 2014; Chen et al., 2020). Additionally, craniofacial characteristics, such as a more retruded position of the upper and lower jaws, a narrower airway, and a shorter soft palate, have also been linked to improved outcomes (Chen et al., 2020).

Nonetheless, none of these factors, whether considered individually or in combination,

provides a consistently reliable prediction. Sutherland et al. (2014) demonstrated in a cohort of 425 MAS-treated patients that standard baseline measures, including demographic, anthropometric, and polysomnography variables, were only weakly correlated with treatment efficacy, highlighting the limitations of current clinical predictors. The authors concluded that alternative objective prediction methods are needed to reliably select patients for the therapy.

OSA is a multifactorial condition, and MAS efficacy may depend on specific pathophysiological traits, or *endotypes*, including upper-airway collapsibility, muscle compensation, arousal threshold, and ventilatory control stability. The study of Edwards et al. (2016) has shown that patients responding to MAS tend to have less collapsible airways, milder anatomical obstruction, and more stable breathing control than non-responders. Although directly measuring these traits requires specialized laboratory techniques not routinely available in clinical practice, researchers have attempted to estimate them from standard polysomnography (PSG) data. For instance, Bamagoos et al. (2019) developed a method to infer OSA characteristics from sleep study recordings, finding that a favorable combination of non-anatomical traits, moderate airway collapsibility, and a less pronounced compensatory muscle response was linked to higher oral appliance efficacy. However, as the authors note, despite promising results, these models still require further validation and refinement.

Anatomical measurements have also been explored as predictors of MAS success. Cephalometric radiographs, in particular, have been examined to identify craniofacial features that correlate with treatment outcomes. A systematic review (Guarda-Nardini et al., 2015) analyzed 13 cephalometric studies and noted that a lower mandibular plane angle and a shorter distance from the hyoid bone to the mandibular plane were frequently observed in responders. Nevertheless, the results across studies have been inconsistent, limiting the reliability of anatomical predictors when used in isolation (Guarda-Nardini et al., 2015).

Advanced imaging techniques such as magnetic resonance imaging (MRI) and computed tomography have provided further insights into the structural and functional changes during mandibular advancement. Jugé et al. (2022) employed dynamic MRI to assess tongue and airway movements during mandibular advancement, reporting that responders showed greater anterior tongue displacement and a larger increase in the oropharyngeal airway cross-sectional area per millimeter of jaw movement compared to non-responders. Similarly, Brown et al. (2021) identified the pterygomandibular raphe as a potential anatomical factor influencing treatment efficacy; patients without this tendon achieved greater mandibular protrusion and, consequently, better outcomes.

More recently, researchers have begun using machine learning techniques to predict

MAS outcomes based on data derived from routine sleep studies. Vena et al. (2020) analyzed nasal airflow signals from diagnostic PSG recordings and extracted several time-domain features, such as the average depth of flow limitation during respiratory events and the shape of the inspiratory and expiratory curves. One particularly noteworthy feature was the "Expiratory Pinching" metric, indicative of soft-palate prolapse during expiration. When combined with clinical variables like age and BMI, this approach yielded a classification accuracy of approximately 74 % in distinguishing responders from non-responders.

Despite these promising results, the study has several methodological shortcomings. Each breath was manually classified for palatal prolapse using video endoscopy during natural sleep: a process that may introduce subjectivity. The Expiratory Pinching metric was calculated for each breath, and the dataset was split evenly into training and test sets to determine an optimal threshold by maximizing the sum of sensitivity and specificity from the Receiver Operating Characteristic curve. This fixed threshold, optimized on the training data, risks overfitting and potential information leakage. Ideally, threshold selection should be performed on an entirely independent dataset to enhance the generalizability of the findings.

Another signal-based approach by Zeng et al. (2007) used flow-volume curves from spirometry of awake patients performed in different postures as a surrogate measure of upper-airway collapsibility. Their findings suggested that responders exhibited a distinct spirometric pattern characterized by a lower mid-inspiratory flow and a higher ratio of expiratory-to-inspiratory flow. Nonetheless, the use of spirometry in OSA workups is limited by its infrequent application in standard clinical protocols.

Signal-based prediction methods offer a promising, non-invasive means of assessing outcomes of treatment with MAS, using data that are already routinely collected during OSA diagnosis. Machine learning methods can enable the identification of complex patterns that might otherwise be overlooked. However, most studies to date have been limited by small sample sizes and a narrow focus on specific signal features. Notably, frequency-domain analyses, which could reveal distinct oscillatory patterns associated with upper-airway dynamics during flow-limited breathing, remain underexplored and may provide additional predictive insights.

### I.2.3 Limitations of Existing Approaches

Despite the progress in this field, no current method offers a fully reliable and clinically practical predictor of MAS treatment outcome. Physiological trait-based approaches, while offering valuable mechanistic insights, often rely on measurements that are not part of routine clinical practice. Similarly, anatomical and phenotypic predictors, although indicative of certain trends, frequently exhibit considerable overlap between responders and

non-responders, limiting their predictive accuracy. Conventional statistical models that integrate demographic, clinical, and signal-based variables have achieved only moderate success. While each approach contributes valuable insights, there remains a significant unmet need for a robust, non-invasive tool to reliably predict efficacy of MAS.

### I.2.4 Research Question

Considering the limitations discussed earlier, this thesis employs a machine learning approach that uses features extracted exclusively from the nasal airflow signal recorded during diagnostic PSG. The central hypothesis is that the airflow rate waveform contains essential information about a patient's upper-airway dynamics, which can predict their response to MAS therapy. Specifically, the study investigates whether it is feasible to accurately classify patients as responders or non-responders to treatment with MAS using only these signal-derived features, thereby eliminating the need for additional invasive or specialized physiological measurements. Addressing this question will deepen our understanding of signal-based phenotyping in OSA and may provide a practical, non-invasive predictive tool for supporting personalized treatment strategies in sleep apnea management.

Chapter II

# Biomedical Signal Sources: Physiological Context and Clinical Measurement

## II.1 Skeletal Muscle Physiology and Relevant Pathophysiology

Skeletal muscle serves as the primary actuator for voluntary movement, enabling essential functions such as posture maintenance, locomotion, and breathing. Muscle contraction, the fundamental mechanism underlying these actions, begins when a nerve cell (motor neuron) releases the neurotransmitter acetylcholine onto the muscle fiber membrane. This chemical interaction is converted into an electrical signal known as an action potential, which propagates rapidly across the muscle fiber membrane (sarcolemma) and down specialized membrane tunnels called transverse-tubules (T-tubules). The action potential triggers the release of calcium ions ($Ca^{2+}$) from an internal storage structure called the sarcoplasmic reticulum. These calcium ions initiate the interaction between actin and myosin filaments within muscle fibers, causing them to slide past each other, shorten the muscle fiber, and produce mechanical force (Pham et al., 2020). Uniform electromechanical coupling across fibers ensures that neural commands translate into precisely timed mechanical work.

The stability and effectiveness of muscle contraction rely significantly on the proper functioning of membrane-bound ion channels. In skeletal muscle, voltage-gated sodium channels, specifically Nav1.4, produce the initial inward current necessary for membrane depolarization, while chloride channels, ClC-1, provide a balancing outward current that stabilizes the membrane potential, facilitating muscle relaxation and preventing unintended repeated activation (Jurkat-Rott et al., 2002; Cannon, 2015). Genetic

mutations affecting these critical channels, notably in genes such as SCN4A (encoding Nav1.4) and CLCN1 (encoding ClC-1), disrupt the precise electrical control of muscle excitability. This disruption leads to a clinical condition called *myotonia*, characterized by delayed relaxation and sustained muscle stiffness following voluntary contraction (Trip et al., 2008; Mitrović et al., 1995).

Collectively, disorders arising from dysfunctions in these ion channels, termed *channelopathies*, illustrate how molecular defects can impair muscle function, resulting in muscle stiffness, episodic weakness, or even transient paralysis. Understanding the structural organization of the fiber, the biophysics of its ion channels, and the resulting patterns of excitability therefore provides an essential foundation for interpreting the abnormal electrical characteristics of the muscle fiber.

### II.1.1 Basic Architecture and Action Potential Generation

Skeletal muscle fibers are elongated, multinucleated cells organized into bundles, known as fascicles, forming the muscle. Each fiber contains parallel structures called myofibrils, composed of repeating units – sarcomeres, giving skeletal muscle its characteristic striated appearance. Sarcomeres contain overlapping thin (actin) and thick (myosin) filaments, whose interaction leads to contraction. The muscle fiber membrane (sarcolemma) surrounds myofibrils and extends into the cell interior as T-tubules, rapidly conducting electrical impulses deep into the fiber (Pham et al., 2020).

Motor neurons connect to muscle fibers at specialized junctions, releasing acetylcholine when activated. This release generates an electrical potential that, upon reaching a certain threshold, initiates a muscle fiber action potential. The action potential travels along the sarcolemma and through the T-tubules, prompting calcium release from the sarcoplasmic reticulum, thus triggering actin-myosin filament interaction and contraction. Each muscle fiber action potential directly precedes and determines the mechanical response, or twitch, of that fiber (Pham et al., 2020).

### II.1.2 Ion Channels and Muscle Fiber Electrophysiology

The electrical excitability of skeletal muscle fibers relies critically on specialized membrane proteins known as ion channels, which regulate changes in membrane voltage. At rest, the muscle fiber maintains a transmembrane potential around –90 mV, primarily due to the activity of potassium channels and sodium-potassium pumps ($Na^+/K^+$-ATPase) (Hille, 1978). Notably, skeletal muscle fibers possess a extraordinarily high resting chloride conductance, mediated by ClC-1 chloride channels, contributing about 80% of the total membrane conductance (Bretag, 1987). This large chloride conductance stabilizes the membrane potential and opposes small depolarizations.

When an action potential is initiated, voltage-gated sodium channels open rapidly, causing

a swift influx of $Na^+$ ions into the muscle fiber and driving a sharp rise (depolarization) in membrane potential. These sodium channels quickly become inactivated, stopping the influx within milliseconds. Subsequently, voltage-gated potassium channels open, allowing $K^+$ ions to flow outward, repolarizing the membrane. Simultaneously, ClC-1 chloride channels, whose opening probability increases with depolarisation, allow $Cl^-$ ions to flow into the fiber, registered electrophysiologically as an outward membrane current. This current helps rapidly restore the resting membrane potential and ensures the muscle fiber quickly returns to its electrically stable state (Adrian et al., 1974).

Under physiological conditions, each nerve impulse typically triggers a single muscle fiber action potential followed by complete repolarization. The high chloride permeability in muscle fibers also serves to counteract the depolarizing after-effects of high-frequency firing: during repeated or prolonged activity $K^+$ ions can accumulate in the T-tubules, which would depolarize the fiber (Almers, 1972). The large chloride conductance provided by ClC-1 channels counteracts these effects, buffering the buildup of depolarizing charges and preventing inappropriate repeated excitation (Tang et al., 2011). Thus, ion channels function together to precisely control muscle excitability: sodium channels initiate contraction, whereas potassium and chloride channels restore stability, ensuring that muscle fibers contract only when appropriately stimulated by the motor nerve.

### II.1.3 Skeletal Muscle Disorders: Skeletal Muscle Channelopathies

Inherited mutations in skeletal-muscle ion-channel genes can derail the normal excitation–contraction sequence, producing a spectrum of diseases collectively called *skeletal muscle channelopathies* (Jurkat-Rott et al., 2002). Two broad electrophysiological syndromes are recognised: fiber *hyperexcitability*, which yields delayed relaxation or stiffness – myotonia, and fiber *hypoexcitability*, which produces episodic paralysis or weakness (Jurkat-Rott et al., 2002). The non-dystrophic myotonias comprise disorders with prominent myotonia but minimal fixed weakness. The primary genetic entities within this group are chloride-channel and sodium-channel myotonias.

Chloride-channel myotonia or myotonia congenita results from loss-of-function mutations in the CLCN1 gene, encoding the ClC-1 chloride channel. ClC-1 contributes substantially to the resting membrane conductance of skeletal muscle fibers. Its dysfunction increases sarcolemmal input resistance, thereby lowering the threshold for depolarization and promoting repetitive firing (myotonia) (Tang et al., 2011). Clinically, patients typically show pronounced stiffness upon initiating movements, which notably improves after repeated muscle contractions: a phenomenon termed the "warm-up" effect. Cold exposure sometimes exacerbates muscle stiffness (Deymeer et al., 1998) and is occasionally followed by transient weakness after strenuous activity. Painful muscle cramps are relatively uncommon. Although both autosomal recessive (Becker type) and dominant (Thomsen

type) forms share similar physiological profiles, Becker myotonia typically manifests later in childhood or adolescence.

Sodium-channel myotonias, caused by mutations in SCN4A, encoding the Nav1.4 sodium channel, exhibit distinct pathophysiological mechanisms. Mutations typically disrupt fast channel inactivation or permit abnormal late reopening, generating sustained inward sodium currents (Cannon, 1997). Moderate disruption leads to repetitive firing and myotonia, whereas more severe impairments result in sustained depolarization, causing temporary paralysis (periodic paralysis) (Cannon et al., 1993). Clinically, sodium-channel myotonias range from paramyotonia congenita, with paradoxically worsening stiffness upon repetitive activity or exposure to cold, to potassium-aggravated myotonia and hyperkalemic periodic paralysis. Painful muscle cramps are common, and cold exposure may precipitate stiffness or flaccid weakness. Onset is typically in early childhood, around four to five years of age.

Chloride-channel and sodium-channel myotonias share muscle stiffness, but there are important differences. A comparative study of Trip et al. (2009) involving sixty-two genetically confirmed patients demonstrated clear clinical differences: the warm-up phenomenon was universally present in chloride-channel myotonia patients but observed in fewer than half of sodium-channel cases. Conversely, paradoxical myotonia and cold-induced weakness were observed exclusively in patients with sodium-channel mutations. Additionally, transient post-exercise weakness was common in chloride-channel myotonia, while painful cramps predominated in sodium-channel myotonias (Trip et al., 2009; Fournier et al., 2006). Despite these differences, clinical phenotype may overlap, which makes additional electrophysiological assessments and genetic analysis necessary.

First-line therapy for non-dystrophic myotonias is typically empirical and aimed at reducing muscle membrane excitability. Mexiletine, a sodium channel blocker and antiarrhythmic drug, has demonstrated efficacy in alleviating repetitive muscle stiffness and is regarded as the treatment of choice (Statland et al., 2012). Alternative medications, including lamotrigine, ranolazine, flecainide, and carbamazepine, may be considered if mexiletine proves ineffective or is poorly tolerated. In patients with sodium-channel mutations associated with episodic paralysis, it is essential to avoid triggers such as cold exposure, prolonged physical exertion, and hyperkalemia. Additionally, carbonic anhydrase inhibitors like acetazolamide or dichlorphenamide can effectively reduce the frequency and severity of paralytic episodes.

## II.2 Electromyography

In these section we will review the basics of needle or intramuscular EMG (iEMG) and discuss the waveforms , which are relevant in the context of this thesis. One of the goals of

this work is to test the universal applicability of the classification algorithm: the first step for this will be to take another waveform, but still recorded by iEMG and see whether it accomplishes the classification task for that waveform as well.
In the context of this thesis, fibrillation potentials serve as a point of contrast to myotonic discharges: both are spontaneous, but fibrillations indicate a loss of neuronal control (e.g. in neuropathic or myopathic disease), whereas myotonic discharges indicate an intrinsic muscle membrane hyperexcitability (channelopathy). The successful classification of other types of waveforms would improve our ability to interpret EMG signals for clinical diagnosis of neuromuscular disorders.

### II.2.1 Fundamentals of Electromyographic Measurements

EMG records the bio-electric activity of skeletal muscles by detecting the extracellular voltage fields produced when trans-membrane currents from active muscle fibers spread through the surrounding tissue. In clinical intramuscular EMG (iEMG), concentric or monopolar needle electrodes are typically inserted into muscle tissue, positioned within a few hundred micrometers of active fibers. These electrodes detect signals in the $\sim 50$ $\mu$V to 5 mV range, whose amplitude and waveform characteristics directly reflect the muscle's structural and physiological state (Kimura et al., 2025).

To translate these signals into usable waveforms, the raw electrode output is initially amplified with a gain factor typically between 1,000 and 10,000 (Tankisi et al., 2020). The amplified signal is then passed through a band-pass filter, removing baseline drift, motion artifacts, and high-frequency electronic noise, while retaining frequencies relevant for diagnostic interpretation, typically from 10–20 Hz up to 10 kHz (Tankisi et al., 2020).

Once analogue conditioning is completed, the signal undergoes digitization. While Nyquist's sampling theorem requires a sampling rate at least twice the maximum frequency present, EMG equipment typically uses a higher sampling rate, about three times the highest frequency component, to provide a safety margin against aliasing. Consequently, modern EMG systems commonly sample at frequencies between 20–50 kHz (Nilsson et al., 1993). The digitised EMG data are then stored for quantitative analysis and simultaneously presented visually on an oscilloscope and audibly via a loudspeaker.

EMG examination involves three distinct physiological stages that together form a comprehensive assessment of neuromuscular function. Initially, the muscle is evaluated at complete rest, where a healthy muscle remains electrically silent except for brief insertional activity occurring as the needle electrode is introduced. Persistent spontaneous electrical activity at this stage, such as fibrillation potentials, positive sharp waves, or high-frequency waxing-waning myotonic discharges, clearly indicates pathology (Stålberg et al., 2019).

Subsequently, during slight voluntary contraction, individual motor-unit action potentials (MUAPs) become evident. Analysis of their amplitude, duration, number of phases, and firing regularity allows differentiation between myopathic and neurogenic disorders. Variations in these parameters reflect alterations in the number and size of muscle fibers within each MU.

Finally, as muscle contraction intensity progressively increases towards maximal effort, the EMG recording transitions into a dense interference pattern. The fullness and stability of this pattern provide insights into the number of recruitable MUs and the efficiency of their activation (Stålberg et al., 2019). Throughout all stages, the amplitude and frequency characteristics of EMG signals depend on the anatomical location assessed, as well as on the underlying cause and type of pathology (Subasi, 2019).

### II.2.2 Fibrillation Potentials

Spontaneous or needle-induced electrical signals in a resting muscle provide critical information about the nature, severity, and duration of an underlying neuromuscular disease. The brief burst that accompanies needle movement – insertional activity, is produced by mechanical disturbance of the muscle-fiber membrane and normally subsides within $\sim 300\,\text{ms}$; persistence beyond this duration is abnormal and is termed spontaneous activity (Stålberg et al., 2019). Fibrillation potentials are one form of such activity measured by EMG, most commonly indicative of denervation.

A fibrillation potential is a spontaneous action potential generated by a single skeletal-muscle fiber in the absence of motor-axon input. It emerges after axonal disruption in conditions such as axonal neuropathies, radiculopathies and motor-neuron disease, and is also seen when destructive myopathies disconnect the fiber from its neuromuscular end-plate (Kimura et al., 2025; Rubin, 2019). Denervated fibers become electrically unstable: the resting membrane potential depolarises by about 10–15 mV (Albuquerque et al., 1968), sodium-channel density increases (Pappone, 1980), chloride conductance falls (Bretag, 1987), and a slow depolarising after-potential develops (Thesleff et al., 1975; Midrio, 2006). Intracellular microelectrode and concentric-needle EMG studies show that each action potential is followed by a stereotyped after-hyperpolarisation and a slow depolarising pre-potential; this oscillation repeats until full repolarisation, accounting for the repetitive, but not perfectly clock-like pattern, which was first analyzed by Buchthal et al. (1966).

Fibrillation activity typically begins to appear a few days to a few weeks, approximately 1–4 weeks, after an acute nerve injury (Willmott et al., 2012). The exact delay varies with anatomy: muscles that lie closest to the point of axonal interruption (paraspinals after a radiculopathy, or proximal limb muscles after a plexus lesion) develop fibrillation activity first, whereas the same nerve's more distal targets may lag by several additional weeks

(Oh, 2003). The presence of fibrillation potentials is the electrophysiological sign of active denervation: in clinical practice, their distribution and density are used to stage the lesion, estimate chronicity, and monitor re-innervation during follow-up (Pond et al., 2014).

On needle electromyography a fibrillation potential is the action potential of a single, denervated muscle fiber. It has a characteristically low voltage and very short duration, and it recurs at a slow, fairly steady rate. Fibrillation potentials are low-amplitude (approximately 10–300 $\mu$V) and brief (1–5 ms) spikes (Preston et al., 2012). They tend to fire in a regular or semi-regular rhythmic pattern, often between 0.5 and 10 Hz, though both slower and faster rates can be observed (Thornton et al., 2012). Immediately after axonal loss the firing pattern is often irregular ("acute" fibrillations); over the ensuing weeks it typically evolves into a more regular, slowly drifting cadence ("chronic" fibrillations). This time-dependent change in rhythm was first analyzed in detail by Buchthal et al. (1966) and Conrad et al. (1972), who showed that the transition parallels progressive depolarization of the resting membrane potential and altered ion-channel expression in the denervated fiber. In practice, when a relaxed muscle is probed and the loud-speaker emits a steady "thrum" of identical low-voltage spikes every few hundred milliseconds, one is hearing fibrillation potentials.

Differentiating fibrillation potentials from voluntary activity can be challenging in certain scenarios. Misinterpretation is most likely when the patient is not fully relaxed: the earliest-recruited MUs in a barely perceptible voluntary effort fire irregularly at about 5–10 Hz and can mimic fibrillation potentials in both rate and morphology (Eberstein et al., 1968). Overlapping firing patterns can also confuse interpretation: in a muscle with many denervated fibers, multiple fibrillation potentials and positive sharp waves may be active simultaneously. Their superimposition can create a low-amplitude interference pattern that superficially resembles a weak voluntary contraction (Kimura et al., 2025). This is further complicated if fasciculation potentials, that is spontaneous MU discharges, are also present, adding larger-amplitude random MUAPs into the mix.

### II.2.3 Myotonic Discharges

Myotonic discharges are another form of spontaneous EMG activity, which arise from muscle membrane hyperexcitability. They typically present as high-frequency bursts of action potentials from individual muscle fibers, characterized by rhythmic waxing and waning in both amplitude and frequency. Myotonic discharges are commonly initiated by needle movement or brief voluntary activation, after which the muscle fibers continue firing spontaneously. Initially firing at high frequencies (about 100–150 Hz and occasionally higher), the discharge rate then waxes and wanes until it falls toward 20 Hz, with a parallel modulation of spike amplitude (Ferrante, 2018). This characteristic activity produces an acoustic signature often described as a "dive-bomber" sound, descending in pitch.

Morphologically, the spikes resemble small fibrillation potentials or positive waves, but the rhythmic frequency-and-amplitude modulation distinguishes myotonic discharges (Preston et al., 2002). Myotonic discharges are found across the nondystrophic myotonias – myotonia congenita, paramyotonia congenita, sodium-channel myotonia, as well as in myotonic dystrophies and certain toxic or inflammatory myopathies. Their presence indicates an abnormal tendency for muscle fibers to continue uncontrolled repetitive firing after initial activation.

The pathophysiological mechanism of myotonic discharges differs depending on whether the underlying disorder is a chloride or sodium channelopathy. Fig. II.1 illustrates the pathophysiology of myotonic discharges in skeletal muscle from the initiation of an action potential in a spinal neuron to the muscle fiber's attempt to relax. The process involves several crucial steps (the red arrows in the figure highlight the delayed relaxation and repetitive discharges, emphasizing the key ionic defects driving myotonia: persistent sodium influx or inadequate chloride-mediated repolarization):

1. **Spinal Neuron → Axon**: An action potential is generated in the spinal neuron and travels along the axon toward the nerve terminal.

2. **Nerve Terminal → Synapse**: Upon arrival at the nerve terminal, voltage-gated calcium ($Ca^{2+}$) channels open, allowing $Ca^{2+}$ influx, which triggers the release of acetylcholine (ACh) into the synaptic cleft.

3. **Synapse → Muscle Fiber**: ACh binds to nicotinic receptors on the muscle fiber membrane, causing sodium ($Na^{+}$) to enter the muscle cell.

4. **Muscle Fiber → T-Tubules**: The action potential propagates along the muscle fiber and down the T-tubules, where voltage sensors activate calcium release from the sarcoplasmic reticulum (SR).

5. **Sarcoplasmic Reticulum, $Ca^{2+}$ Ion Channels**: Released $Ca^{2+}$ interacts with the contractile apparatus, leading to muscle contraction.

6. **Ion Channels (Normal vs. Myotonic)**: In normal muscle relaxation, chloride ($Cl^{-}$) channels open, and sodium channels close to repolarize the muscle fiber. In myotonia: some fraction of $Na^{+}$ channels do not fully inactivate and continue to depolarize the membrane or even reactivate, resulting in a chain of self-sustained action potentials (Mitrović et al., 1995; Cannon et al., 1993). In chloride channel myotonia, the reduced chloride conductance leads to prolonged after-depolarizations. After an initial voluntary contraction, the combination of residual depolarization and accumulated $K^{+}$ in the T-tubules can keep the membrane above its firing threshold, causing the fiber to repetitively discharge until the membrane gradually repolarizes back to rest (Cannon, 1996a).

In both cases, the discharges ultimately stop when ionic gradients return to balance or when the abnormal channels eventually close; however, this process may last several seconds, during which the distinctive waxing-and-waning burst pattern appears on iEMG recordings (Ferrante, 2018).

The clinical behavior of the discharges differs: in myotonia congenita, repeated muscle use tends to shorten or abolish the runs (consistent with the warm-up phenomenon, as muscle fiber excitability normalizes with exercise), whereas in paramyotonia congenita, repetitive activity or cooling can provoke more prolonged or frequent runs (paradoxical myotonia) (Fournier et al., 2006). Cooling of muscle typically exacerbates the myotonic discharges in sodium channel myotonia, EMG after cooling will show longer trains or even a transition into electrically silent weakness once the fiber enters depolarization block (Fournier et al., 2006). In contrast, warming up the muscle through repeated contraction often reduces the occurrence or duration of myotonic bursts in chloride channel myotonia, correlating with the clinical improvement of stiffness (Fournier et al., 2006).

## II.3 Obstructive Sleep Apnea

OSA is a common sleep-related breathing disorder characterized by recurrent episodes of partial or complete upper airway collapse during sleep, despite ongoing respiratory effort (Dempsey et al., 2010; Sutherland et al., 2014). Clinically, OSA manifests as loud snoring, fragmented sleep, and daytime symptoms such as excessive sleepiness. However, the effects of OSA extend far beyond disrupted sleep, as the disorder is associated with a range of adverse health conditions, spanning neuropsychiatric issues and cardiovascular complications, that can undermine work performance, family relationships, and social well-being (Moyer et al., 2001). Even when adjusting for prevalent conditions such as obesity, hypertension, and cardiovascular disease, individuals with OSA demonstrate a pronounced decline in both physical health and mental well-being compared to the general population (Bjornsdottir et al., 2015).

We begin our discussion of OSA by describing the anatomical and physiological features of the upper airway that render it susceptible to collapse. Next, we explore the pathophysiological mechanisms that lead to airway obstruction, with emphasis on neuromuscular control, tissue compliance, and ventilatory stability. The subsequent subsection highlights the use of PSG as a diagnostic tool, explaining how the analysis of airflow signals captured during PSG aids in evaluating OSA severity and uncovering its distinct physiological characteristics. By establishing the link between upper airway anatomy, its dynamic function, and the clinical manifestations of OSA, this section lays the ground for later discussions on targeted therapeutic strategies and their implications for effective management.

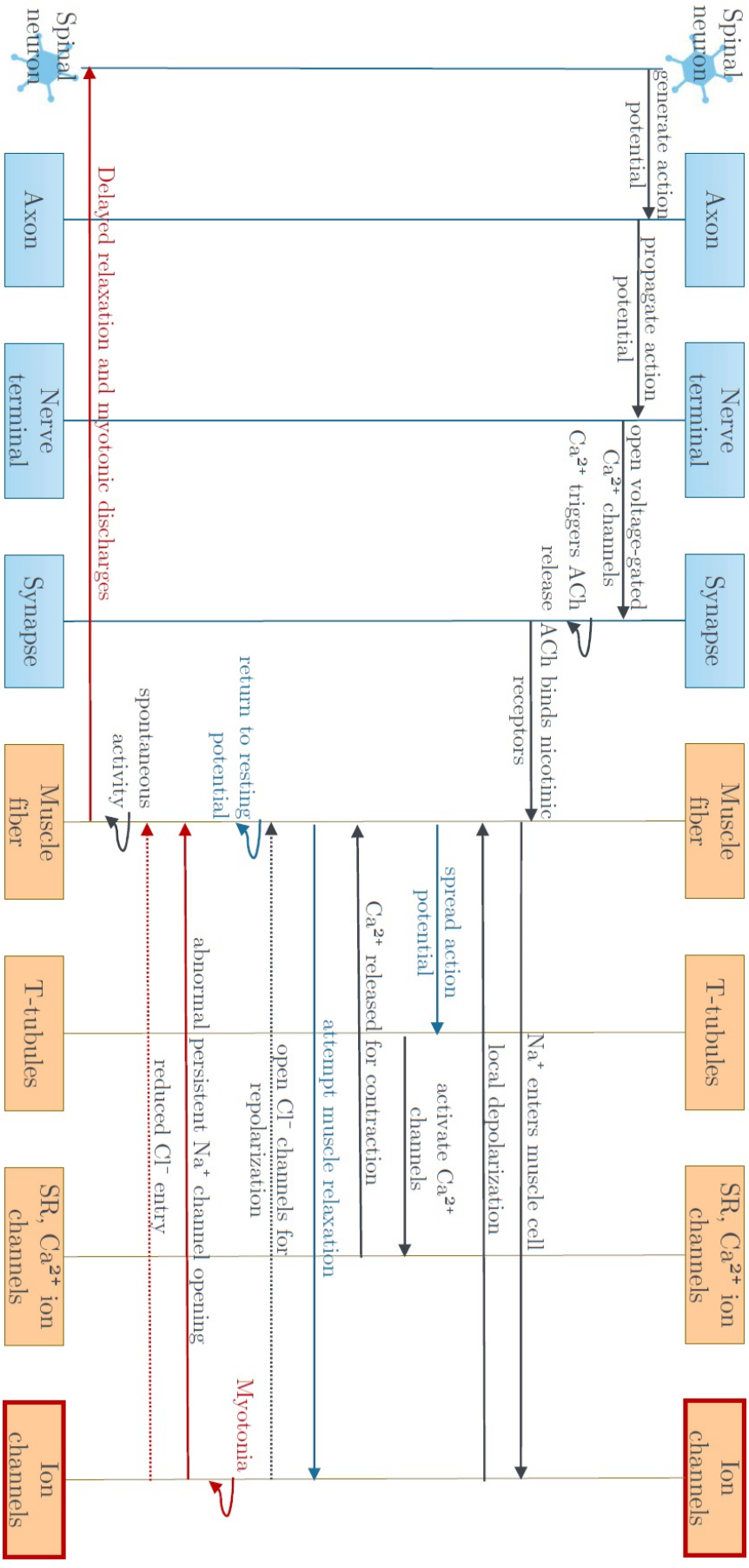



**Figure II.1:** The pathophysiology of myotonic discharges in skeletal muscle.

### II.3.1 Anatomy and Physiology of the Upper Airway Relevant for Development Obstructive Sleep Apnea

The upper airway extends from the nasal openings and mouth to the larynx and is composed of a series of soft, flexible passages, namely, the nasopharynx, oropharynx, and hypopharynx that function as a collapsible tube. Unlike the lower airway (trachea and bronchi), which is supported by rigid cartilage, the upper airway relies on muscles and connective tissues for structural support. Its patency is maintained by a balance between the internal air pressure and the dynamic forces generated by these soft tissues.

During wakefulness, airway patency is actively maintained through neuromuscular compensation. As inhalation occurs, the resulting drop in upper airway pressure initiates a reflexive contraction of throat muscles. This coordinated muscular response stabilizes the soft palate and positions the tongue forward, thereby widening and stiffening the airway lumen. Even in individuals with anatomically narrower airways, the enhanced muscle tone during wakefulness acts as a natural splint, maintaining airway patency. Electromyographic studies confirm that those predisposed to OSA exhibit elevated activity in these supportive muscles while awake, compensating for their structural limitations (Malhotra et al., 2002; Dempsey et al., 2010).

The anatomical characteristics of the upper airway play a critical role in OSA pathogenesis. A narrow or crowded airway operates closer to its collapse threshold, so even a modest reduction in muscular support or a slight increase in negative inspiratory pressure can trigger partial closure. Conversely, an airway with a larger diameter or firmer structural support better resists these forces. Vulnerabilities such as increased soft tissue volume, from fat deposits, enlarged tonsils, or a bulky soft palate, and craniofacial traits like a retruded lower jaw further diminish the available airway space. Consistently, imaging studies reveal that individuals with OSA exhibit a smaller cross-sectional area in key regions of the upper airway along with an increased volume of surrounding soft tissue compared to those without the condition (Dempsey et al., 2010; Malhotra et al., 2002; Isono et al., 1997).

### II.3.2 Pathophysiology of Obstructive Sleep Apnea

OSA is caused by the repetitive collapse of the pharyngeal airway during sleep. As an individual transitions from wakefulness to sleep, especially during non-rapid eye movement (non-REM) sleep, a generalized reduction in both steady and reflexive muscle tone diminishes the neuromuscular support that normally maintains airway patency, predisposing the airway to narrowing and collapse during inspiration.

A central concept in understanding this process is the *critical closing pressure* ($P_{\text{crit}}$). This parameter defines the threshold below which the airway collapses due to insufficient internal pressure. In many patients with OSA, a smaller or more compliant airway results

in an elevated $P_{\text{crit}}$ that may approach or exceed atmospheric pressure. Under these conditions, even modest negative pressures generated during normal inhalation can trigger collapse (Dempsey et al., 2010).

Obstructive events are commonly terminated by an EEG-defined arousal or by a rapid rise in ventilatory drive that restores upper-airway dilator activity (Dempsey et al., 2010; Jordan et al., 2011). Recurrent cycles of obstruction and relief fragment sleep and generate intermittent hypoxemia and hypercapnic excursions together with large negative intrathoracic pressure swings; these perturbations are associated with sympathetic activation and cardiovascular stress (Dempsey et al., 2010; Yeghiazarians et al., 2021). At the population level, OSA is linked to elevated cardiovascular risk, although individual susceptibility varies, consistent with heterogeneity across key physiological traits such as airway collapsibility, upper-airway muscle responsiveness, ventilatory-control instability (loop gain), and arousal threshold (Eckert et al., 2013).

The severity and individual susceptibility to OSA depend on several interrelated factors. Beyond anatomical predisposition, such as a narrow or highly compliant upper airway, the overall severity is influenced by the efficiency of neuromuscular compensation during sleep (Dempsey et al., 2010). Another critical factor influencing OSA severity is the stability of the respiratory control system, characterized by a parameter known as loop gain. Loop gain quantifies the sensitivity of the respiratory system's response to disturbances; a high loop gain indicates that even minor changes in blood oxygen or carbon dioxide levels elicit exaggerated ventilatory reactions, promoting instability and frequent recurrence of airway obstruction. Additionally, the arousal threshold, the level of respiratory disturbance necessary to trigger awakening, significantly affects the frequency and severity of obstructive events. Individuals with a low arousal threshold awaken more readily in response to minor breathing disruptions. While this can prevent prolonged or severe airway obstruction, it often results in substantial sleep fragmentation.

In summary, OSA arises from the interaction of anatomical collapsibility ($P_{\text{crit}}$), neuromuscular responsiveness, ventilatory-control stability (loop gain), and arousal threshold. This four-trait framework helps explain individual variability but has limited clinical application, as targeting single traits rarely yields consistent therapeutic benefit and noninvasive endotype estimates show moderate night-to-night variability. Thus, while endotyping remains valuable for mechanistic insight and research, current treatment strategies continue to rely mainly on established interventions such as CPAP, mandibular advancement, and positional or anatomical approaches (Eckert et al., 2013; Finnsson et al., 2023; Magalang et al., 2022).

### II.3.3 Polysomnography Overview and Its Role in Obstructive Sleep Apnea Diagnosis

PSG is a comprehensive overnight sleep study that is considered the gold standard for diagnosing OSA in adults (Kapur et al., 2017). In-laboratory PSG simultaneously records multiple physiological signals during sleep, allowing clinicians to monitor sleep stages and detect the characteristic breathing abnormalities of OSA. Performed in accredited sleep centers under the supervision of trained technologists and sleep physicians, PSG ensures high-quality, interpretable data.

In a typical PSG procedure, the patient arrives at the sleep laboratory in the evening for an overnight study. Following a brief orientation to the procedure, surface electrodes and sensors are applied to record various physiological signals. Electroencephalography (EEG) electrodes placed on the scalp monitor brain activity, while electrooculography (EOG) sensors near the eyes track eye movements. EMG sensors positioned on the chin and legs record muscle tone and limb movements, and electrocardiography (ECG) leads monitor heart rhythm. Respiratory signals are captured using a nasal pressure cannula or an oronasal thermistor to measure airflow at the nose and mouth, supplemented by elastic belts placed around the chest and abdomen to assess respiratory effort. Additionally, a finger pulse oximeter continuously tracks blood oxygen saturation throughout the night.

After sensor placement, the patient is encouraged to sleep naturally, and lights are typically turned off at their habitual bedtime. The PSG recording then continues uninterrupted for approximately 6–8 hours. Patients can freely move and change positions during sleep despite the attached equipment. Throughout the recording period, a technologist continuously monitors data from an adjacent control room to note significant respiratory or sleep events and ensure optimal signal quality. In certain cases, such as when severe OSA is identified early during the recording, the study may transition into a "split-night" protocol. In this scenario, the first portion of the night is used for diagnostic assessment, and if sufficient evidence of OSA is established, the second part involves positive airway pressure titration to promptly initiate therapy. Upon completion of the study, a sleep specialist analyzes and scores the PSG data according to standardized criteria.

PSG data enable the identification and quantification of respiratory events characteristic of OSA, namely apneas and hypopneas. In adults, an *apnea* is scored when airflow decreases by $\geq 90\%$ from baseline for at least 10 s. A *hypopnea* is scored when the peak nasal pressure (or PAP flow) signal decreases by $\geq 30\%$ for $\geq 10$ s and is associated with either a $\geq 3\%$ oxygen desaturation or an arousal; an acceptable alternative uses a $\geq 4\%$ desaturation without arousal (American Academy of Sleep Medicine, 2023; Berry et al., 2012; American Academy of Sleep Medicine, 2013).

During obstructive respiratory events, airflow ceases or diminishes due to physical

narrowing or collapse of the upper airway, despite continued respiratory effort. This ongoing respiratory effort is typically visible in PSG through thoracic and abdominal movement channels, sometimes appearing as paradoxical movements in which the chest and abdomen move in opposite directions. The persistent breathing effort clearly distinguishes obstructive events from central apneas, during which airflow cessation occurs without corresponding respiratory muscle activity. Most obstructive apneas or hypopneas terminate with an arousal, visible on EEG as a transient burst of faster brain-wave activity lasting approximately 3–15 seconds. These brief awakenings represent the brain's response to changes in blood gas levels and the strain of attempting to breathe against an obstructed airway.

To ensure consistent and reproducible diagnosis, respiratory events identified in PSG are scored according to standardized guidelines developed by the American Academy of Sleep Medicine. The total number of apneas and hypopneas observed during the sleep period is calculated and reported as the *apnea-hypopnea index* (AHI), expressed in events per hour of sleep. The AHI serves as the principal clinical metric for diagnosing OSA and evaluating its severity (Kapur et al., 2017). An AHI of 5 or more events per hour, accompanied by characteristic symptoms such as daytime sleepiness, confirms an OSA diagnosis in adults. Increasing AHI values correspond to greater disease severity, with AHI thresholds of 15 and 30 events per hour generally indicating moderate and severe OSA, respectively (Kapur et al., 2017). Ultimately, by recording the frequency and duration of apneas and hypopneas, along with the corresponding oxygen desaturations and sleep disruptions, PSG enables a definitive diagnosis of OSA and informs subsequent management strategies.

OSA arises from the vulnerability of the upper airway, which, due to specific anatomical and functional characteristics, fails to remain open when muscle tone naturally decreases during sleep. This leads to airway collapses that produce distinctive alterations in airflow signals and trigger a cascade of physiological disturbances. Analysis of airflow signals captured during PSG can offer insights into airway dynamics and the underlying mechanisms of obstruction. In subsequent section, we will discuss management strategies for OSA and examine the anatomical and physiological factors that influence treatment outcomes. Using PSG data can improve our ability to predict therapeutic responses, identify patients most likely to benefit from treatment, and quantify physiological improvements, thereby supporting more personalized and effective OSA management.

## II.4 Mandibular Advancement Splints

In adults, *positive airway pressure* (PAP) – most commonly *continuous* PAP (CPAP), remains the first-line therapy for OSA. PAP delivers a constant supra-atmospheric pressure through a nasal or oronasal interface, pneumatically splinting the upper airway and preventing collapse during sleep. Pressure levels are determined either by attended

titration or through auto-adjusting PAP, while bilevel PAP can be used when higher pressures or intolerance to fixed CPAP occur (Patil et al., 2019).

Although CPAP is highly effective physiologically, its real-world impact is limited by variable adherence. Common reasons for discontinuation include mask discomfort and leakage, nasal or oral dryness, aerophagia, device noise, bed-partner disturbance, claustrophobia, and pressure intolerance. Early nightly use strongly predicts long-term compliance (Weaver et al., 2008; Rotenberg et al., 2016). Mask design also influences tolerance: oronasal masks typically require higher pressures, are more leak-prone, and are associated with poorer adherence than nasal interfaces; addressing nasal symptoms and favoring nasal masks when possible can therefore improve outcomes (Genta et al., 2020; , 2023; Rotty et al., 2021). Patient- and disease-related factors such as symptom severity, smoking, nasal obstruction further affect usage.

Because of these limitations, alternative non-PAP treatments are important for patients who cannot tolerate or decline PAP therapy. One established option is the *mandibular advancement splint* (MAS) – a custom, titratable oral appliance worn during sleep that protrudes the mandible to enlarge the velopharyngeal airway and reduce its collapsibility. These devices are fabricated from dental impressions or digital scans taken by professionals, ensuring a good fit and enhanced comfort (Manetta et al., 2022). Modern MAS are constructed using various biocompatible materials optimized for intraoral use, and treatment protocols typically prescribe mandibular advancement at 50 % to 80 % of a patient's maximal protrusion, with the primary clinical goals being the reduction of obstructive events, sleep fragmentation, and oxygen desaturation (Basyuni et al., 2018).

In this section, we first review how MAS physically modify upper airway anatomy to reduce its collapsibility. We then examine how these anatomical changes are reflected in polysomnographic signals, particularly the airflow rate waveform, which captures breathing dynamics and patterns of obstruction. This overview provides the base for understanding both the mechanisms of MAS and the role of airflow signal analysis in predicting treatment outcomes for patients with OSA.

### II.4.1 Treatment with Mandibular Advancement Splints in Obstructive Sleep Apnea

MAS treat OSA by holding the lower jaw (mandible) in a forward position relative to the upper jaw (maxilla). This anterior repositioning shifts adjacent soft tissues, particularly the tongue and soft palate, forward, resulting in an enlarged upper airway and reduced airway collapsibility (Mohammadieh et al., 2023). Imaging studies have demonstrated that mandibular advancement increases the velopharyngeal airway space, often by inducing lateral expansion of the pharynx through tension in the soft tissue connections between the mandibular ramus and the lateral pharyngeal walls (Ryan et al., 1999). Additionally,

MAS typically advance the base of the tongue forward because the genioglossus muscle attaches directly to the mandible, thus increasing oropharyngeal space. Dynamic MRI studies confirm that the use of oral appliances results in anterior displacement of the tongue (Brown et al., 2013) and increases the distance between the tongue and soft palate (Kato et al., 2000).

Under passive conditions, such as during muscle atonia in sleep, mandibular advancement markedly reduces pharyngeal collapsibility by lowering the critical closing pressure ($P_{\text{crit}}$). In experiments conducted under anesthesia and paralysis, mandibular advancement increased the cross-sectional areas of both the velopharynx and oropharynx and reduced $P_{\text{crit}}$, thereby stabilizing the airway (Isono et al., 1995). Consistent with this passive mechanism, Edwards et al. (2016) demonstrated in a detailed physiological study that oral appliance therapy improves upper-airway collapsibility under both passive and active conditions, without altering loop gain, arousal threshold, or pharyngeal dilator muscle responsiveness. Importantly, patients with milder baseline collapsibility and lower loop gain experienced the greatest reductions in AHI, highlighting that anatomical factors and ventilatory-control stability strongly influence treatment response.

Complementary findings from Bamagoos et al. (2019) further support this view. Their polysomnographic analysis showed that MAS treatment is most effective in patients with lower pharyngeal muscle compensation, indicating that weaker dilator muscle activity is associated with better outcomes. This suggests that the principal mode of MAS action is passive anatomical enlargement rather than enhancement of neuromuscular activation. Additional predictors of MAS treatment success identified by Bamagoos et al. (2019) include lower loop gain, higher arousal threshold, lower ventilatory response to arousal, and moderate pharyngeal collapsibility. In contrast, demographic and clinical characteristics such as younger age, lower body mass index, female sex, milder OSA severity, and supine dependence have shown only limited and inconsistent predictive value (Sutherland et al., 2014).

Because MAS primarily remodel the airway geometry, their impact on conventional metrics, such as the AHI, can be inconsistent. In some patients, MAS markedly reduce the AHI, while in others the AHI remains relatively unchanged despite objectively improved airway patency. This discrepancy arises because the AHI merely quantifies the number of apneas and hypopneas and does not capture improvements in the quality of breathing. For instance, MAS may transform severe apneas into milder, flow-limited hypopneas that produce less oxygen desaturation and fewer arousals. Consequently, the overall airflow becomes smoother and less flattened, indicating reduced upper-airway obstruction. Patients with such improvements often experience better sleep continuity and improved daytime symptoms, even if the residual AHI remains elevated.

Analysis of airflow signal morphology may offer a more sensitive measure of therapeutic benefit than the AHI alone. Mann et al. (2019) demonstrated that quantifiable changes in the inspiratory airflow, such as a reduction in the flow limitation plateau, decreased flattening, and normalization of the inspiratory profile, reflect improvements in airway patency that are not evident from simply counting apneas and hypopneas. By extracting metrics like the degree of flattening and the presence of a concave inspiratory shape, their approach evaluates the magnitude of pharyngeal obstruction on a breath-by-breath basis, yielding a continuous measure of obstruction severity.

It is important to note that MAS do not completely eliminate OSA in all patients; approximately 30–40% of individuals remain significant non-responders (Sutherland et al., 2014; Chen et al., 2020). Nonetheless, even partial responders often experience meaningful improvements, such as reduced snoring, better oxygenation, and less sleep fragmentation. Consequently, clinical management of MAS therapy frequently considers factors beyond the AHI, including subjective sleep quality and comorbid outcomes like blood pressure reduction. Indeed, evidence indicates that MAS therapy can improve blood pressure, reduce daytime somnolence, enhance driving performance, and boost quality of life to an extent comparable to CPAP in adherent patients, despite CPAP's greater impact on the AHI (Cistulli et al., 2004; Phillips et al., 2013). By directly addressing the anatomical causes of OSA and reducing airway collapsibility, MAS offer a viable and more tolerable alternative to CPAP, yielding significant health benefits even when traditional numerical indices do not fully capture treatment success.

### II.4.2 Relevant Polysomnographic Signals and Their Significance for Prediction of Mandibular Advancement Splint Therapy Outcomes

In context of predicting of MAS treatment outcome, among the various PSG signals, the airflow rate waveform is of particular interest, as it contains rich information about upper airway patency and breathing dynamics. Analysis of its morphology and frequency content can reveal patterns of obstruction and flow limitation that go beyond what is reflected by simple event counts like the AHI. This subsection describes the characteristics of the airflow signal and explains their potential relevance for predicting outcomes of MAS treatment in OSA patients.

During unobstructed normal breathing, the inspiratory airflow waveform (Fig. II.2a) has a smooth, sinusoidal or bell-shaped contour. Flow rises rapidly at the start of inspiration, reaches a peak, and then descends smoothly as lung inflation slows toward the end of inspiration; expiration produces a similar mirror-image curve. This roughly symmetric, rounded shape indicates minimal resistance in the upper airway throughout the breath.

In contrast to the smooth, bell-shaped waveform of normal breathing, airflow signals

in OSA often exhibit abnormal patterns (Fig. II.2b–Figure II.2d). For example, the inspiratory flow may become flattened, display a "sawtooth" pattern, or assume an abnormally concave, scooped shape. These deviations indicate inspiratory flow limitation resulting from partial airway collapse, and they can be quantified to reveal underlying physiological mechanisms.

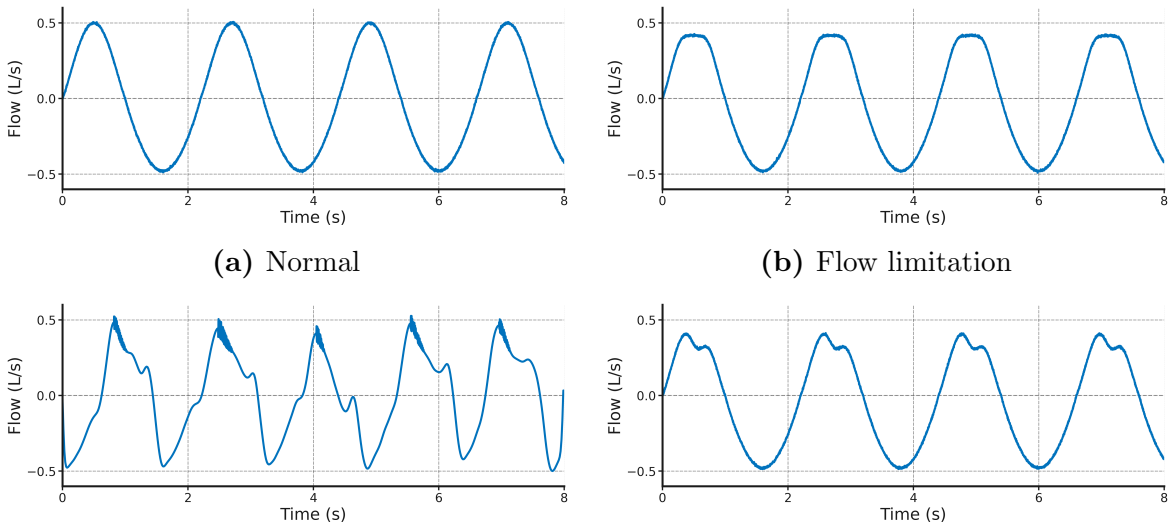
One mechanism is Negative Effort Dependence (NED), which describes the paradoxical decline in airflow despite increasing respiratory effort (Beeck et al., 2024). In normal breathing, inspiratory flow rises rapidly to a peak and then tapers smoothly; however, in flow-limited breaths the inspiratory contour flattens or even down-slopes mid-inspiration (Fig. II.2b), reflecting that the airway has reached its critical collapsibility. This classic sign of flow limitation, characterized by a loss of the normal peak and a prolonged plateau, often defined by a reduction in peak-to-plateau flow (Hosselet et al., 1998; Genta et al., 2017). For instance, collapse at the soft palate typically yields a plateaued flow despite continued respiratory effort (Genta et al., 2017).

In addition to flattening, the airflow signal may exhibit high-frequency oscillations, known as the "sawtooth" pattern (Fig. II.2c). In awake pulmonary function testing, the presence of a sawtooth sign on the inspiratory flow-volume loop is a well-established indicator of upper airway obstruction due to tissue flutter (Levent et al., 2011). Similarly, during overnight PSG, the unfiltered nasal cannula pressure signal can reveal a serrated, oscillatory waveform during inspiration. These intra-breath oscillations result from the vibration of pharyngeal tissues, essentially capturing the snoring phenomenon, which produces the characteristic "sawtooth" appearance in the airflow trace (American Academy of Sleep Medicine, 2010; American Academy of Sleep Medicine, 2008). An automated method for detecting snoring in the airflow signals of OSA patients has been described in Lee et al. (2015).

Another abnormal airflow morphology observed in OSA is "inspiratory concavity", characterized by a pronounced inward dip in the inspiratory flow-time curve (Fig. II.2d). In these cases, the initial portion of inspiration may appear normal, but a significant dip occurs in mid-to-late inspiration, sometimes with multiple inflection points, indicating severe airway collapse. This pattern, reflecting the degree of NED, represents a rapid decline in flow after the initial peak; in extreme cases, it may present as an abrupt drop or discontinuity mid-breath (Beeck et al., 2024). Such abrupt declines have been specifically linked to epiglottic collapse (Genta et al., 2017; Azarbarzin et al., 2017).

Distinct airflow patterns correlate with the anatomical sites of airway collapse. Concurrent sleep endoscopy studies have demonstrated that different collapse sites yield unique flow signal characteristics. For example, (Genta et al., 2017) quantified inspiratory NED and found that tongue-base collapse results in only a slight reduction from peak flow, indicating

minimal NED, whereas isolated palatal or lateral wall collapse is associated with moderate flattening of approximately a 45% reduction in flow. In essence, breaths with tongue obstruction maintain a near-normal inspiratory shape, while those with palatal or lateral wall obstructions display a flattened or snore-oscillatory pattern, and epiglottic collapse produces a markedly concave, jagged profile (Beeck et al., 2024). This relationship between airflow morphology and the site of collapse has been used to phenotype OSA, enabling clinicians to infer the likely anatomical source of obstruction from the airflow contour (Genta et al., 2017; Beeck et al., 2024).




**(a)** Normal **(b)** Flow limitation

**(c)** Sawtooth **(d)** Inspiratory concavity

**Figure II.2: Characteristic inspiratory airflow morphologies observed in polysomnography.** Stylized representations illustrate the major inspiratory waveform patterns: (a) *Normal* — a bell-shaped contour representing unobstructed breathing; (b) *Flow limitation* — a flattened waveform reflecting partial upper airway collapse; (c) *Sawtooth* — high-frequency oscillations superimposed on the inspiratory phase, corresponding to tissue vibration and snoring; (d) *Inspiratory concavity* — a scooped contour, often linked to severe dynamic collapse.

Given that distinct airflow patterns reflect the underlying anatomical characteristics in OSA patients, these signal features may serve as predictors of treatment outcomes with MAS. For example, epiglottic collapse, characterized by a markedly concave inspiratory waveform, has been linked to poorer responses to oral appliance therapy, with endoscopic studies showing that patients with epiglottic obstruction are more likely to experience treatment failure (Kent et al., 2015). Conversely, when the primary site of collapse is at the tongue base or oropharynx, and the soft palate remains relatively stable, mandibular advancement splints tend to yield better outcomes because advancing the mandible directly enlarges and stabilizes the retrolingual airway (Ng et al., 2006).

Linking airflow signal patterns to specific sites of airway collapse can provide valuable insight for predicting the outcomes of MAS therapy. MAS efficacy is highly variable, some patients achieve near-complete resolution of OSA, while others show only minimal improvement (Chen et al., 2020). This variability reflects the heterogeneous nature of OSA, where the relative contributions of anatomical factors and non-anatomical influences, such as neuromuscular compensation and arousal threshold differ among patients. Although the exact mechanisms by which MAS effects remain unclear (Chan et al., 2020; Chen et al., 2020), analyzing airflow signal patterns in baseline PSG may help identify markers that predict a positive response to therapy with MAS.

## Chapter III

# Data and Hypotheses

In this chapter, we introduce the primary datasets employed throughout this thesis. These data sets are used for two distinct investigations: (1) the classification of skeletal muscle disorders using single channel iEMG signals, and (2) the analysis of multi-channel PSG data to predict the outcome of MAS therapy in patients with OSA.

We begin by presenting *skeletal muscle disorders datasets*, which include both clinical and simulated EMG data for studying muscle channelopathies and related myotonic discharges. These data help validate our iEMG classification approach and provide insight into ion channel-specific abnormalities, discussed in the Subsection II.1.3. We also present a separate dataset of fibrillation potentials, which will be leveraged to test the broader applicability of our single-channel signal classification framework.

Subsequently, we will introduce *OSA datasets*, each of which features multiple channels recorded in overnight PSG. Whereas single-channel iEMG dataset captures, only one type of signal, that is muscle electrical activity, PSG data reflect a diverse set of physiological signals, such as respiratory airflow, EEG, EOG, ECG and other types of biomedical signals. This multi-channel nature accommodates a comprehensive view of sleep architecture and breathing patterns, yet also increases the complexity of data handling.

Despite these differences, both the skeletal muscle disorders datasets and the OSA datasets share a common methodological goal: to leverage advanced signal processing and automated classification methods in addressing clinically relevant diagnostic questions. By selecting these complementary datasets, we illustrate how similar machine learning pipelines and feature-engineering principles can be adapted for distinctly different neuromuscular and respiratory conditions.

The remainder of this chapter is structured as follows:

- **Part I (§III.1):** We detail the skeletal muscle disorders datasets, including both measured and simulated data describing skeletal muscle channelopathies (§III.1.1, §III.1.2), as well as a separate fibrillation potentials dataset (§III.1.3). Then we present our hypotheses for each dataset (§III.1.4), outlining the testing strategy for automated classification algorithm.
- **Part II (§III.2):** We introduce OSA datasets and present our hypotheses for predicting OSA treatment outcomes with MAS using respiratory airflow signals (§III.2.1, §III.2.2).

Overall, this chapter functions as a bridge between the clinical and physiological foundations discussed previously and the methodological approaches that will be detailed in subsequent chapters. By the end, we will have laid out the key data resources and hypotheses that underpin our investigations into both skeletal muscle pathologies and OSA treatment efficacy.

## III.1 Skeletal Muscle Disorders Datasets

### III.1.1 Skeletal Muscle Channelopathy Dataset

The skeletal muscle channelopathy data was initially obtained in Radboud University Nijmegen Medical Centre in the time period from April 2005 to March 2006 for the study published at Drost et al. (2015). Patients from across the Netherlands with reported non-dystrophic myotonic syndrome underwent clinical examination, which included iEMG testing and genetic analysis of blood samples. Patients who were found to have myotonic dystrophy types 1 or 2, primary periodic paralysis without myotonic symptoms, or no mutations in the SCNA4 or CLCN1 genes were excluded from the study.

The Medical Ethics Committee of the Radboud University Nijmegen Medical Center approved the study, and all recruited patients provided written consent to participate (Drost et al., 2015). Zoia Lateva, Dick Steegeman, and Kevin McGill, co-authors of Drost et al. (2015), provided us with the dataset used in the present research, from which all personally identifiable information had been removed.

The study by Drost et al. (2015) included a total of 66 participants (35 men and 31 women). This cohort comprised 32 patients with chloride channelopathy, of whom 29 were diagnosed with Becker myotonia congenita and 3 with Thomsen myotonia congenita. The sodium channelopathy group included 22 participants: 5 with paramyotonia congenita and 17 with sodium-channel myotonia. These patient groups, along with their genotypes, were initially reported in Tables 3b and 4 by Trip et al. (2008). The study also included twelve more sodium channelopathy patients, evenly split between paramyotonia congenita (6 patients) and sodium channel myotonia (6 patients), thus bringing the total number

of participants with sodium channelopathy to 34. For a detailed list of these additional patients' diagnoses and genotypes, refer to the supplementary table in the appendix of Drost et al. (2015).

iEMG signals were recorded from five muscles in each patient using a concentric needle electrode, specifically the left orbicularis oculi, left biceps brachii, right first dorsal interosseous, right rectus femoris, and left tibialis anterior. Typically, four recordings were taken from each muscle according to a nerve stimulation protocol comprising four distinct recordings. During signal acquisition in protocols 1 and 4, the examiner gently moved the needle every six seconds to induce myotonic discharges in resting muscles. Recordings obtained in protocols 2 and 3 were performed during slight and maximal voluntary contractions, respectively. The EMG signals were recorded with a sampling frequency of 44,100 Hz and each recording has a duration of approximately 30 seconds. For detailed descriptions of the EMG recording and genetic analysis procedures, see (Drost et al., 2015).

The EMG recordings were initially received in `.wav` format and then converted to `.csv` for further processing. In total, 1,350 recordings were available, with 683 from sodium channelopathy patients and 667 from those with chloride channelopathy. Each file name begins with a lowercase `p` (for 'patient'), followed by the patient's identification number, the abbreviation of the genetic mutation type, the muscle examined, and the protocol number under which the recording was performed. The genetic mutation abbreviations are `CR` (chloride recessive), `CS` (chloride sporadic), `CPR` (chloride probably recessive), `CD` (chloride dominant), and `SD` (sodium dominant). Muscle abbreviations are as follows: left orbicularis oculi (`L_ORB_OCU`), left biceps brachii (`L_BICEP`), right first dorsal interosseous (`RFD_INT`), right rectus femoris (`R_RECT_FEM`), and left tibialis anterior (`L_TIB_ANT`). For example, an EMG recording from the left tibialis anterior of patient 7, diagnosed with a chloride recessive mutation and acquired under protocol 3, would be named: `p07CR_L_TIB_ANT_3.wav`.

To isolate only those recordings containing distinguishable myotonic discharges, we created a table for each patient that listed all recordings and grouped them by muscle. Each table also had columns indicating whether a file contained a clearly detectable myotonic discharge and whether it was severely contaminated by other activities, for example voluntary contraction potentials. After creating these tables, we assessed the EMG recordings in a manner consistent with clinical practice by listening for the characteristic "dive bomber" sound of myotonic discharges. We noted in the table whether a myotonic discharge was audibly identifiable to a trained ear and whether significant contamination might interfere with its detection. Only files with audibly discernible myotonic discharges and no serious contamination were retained in our myotonic discharge database. Consequently, we selected 279 files for the sodium channelopathy group and 221

files for the chloride channelopathy group, recorded from 34 and 32 patients, respectively. As a result, 41% of the files belonged to the chloride channel defect, and 56% to the sodium channel defect group.

It should be noted that the original database did not contain exactly four recordings for every patient and muscle. In some cases, up to seven recordings were made for a single muscle, whereas in others, fewer than four (or none at all) were available, though such instances were rare. Furthermore, the left rectus femoris was examined in patients 15 and 19 instead of the right rectus femoris, and the right biceps brachii was examined in patient 35 instead of the left biceps brachii. The quality of the recordings, in terms of the clarity of audibly distinguishable myotonic discharges and the presence of other activities, varied by patient and muscle. Consequently, more recordings were selected from some patients than from others, and in certain cases no recordings were chosen for a given muscle if they did not meet the selection criteria. Most selected EMG recordings were acquired according to protocols 1 and 4, whereas protocol 3 yielded the fewest recordings included in the final dataset.

### III.1.2 Simulated Skeletal Muscle Channelopathy Dataset

In order to create a simulated skeletal muscle channelopathy database, we used a multi-domain model developed by Klotz et al. (2020). This model includes a detailed biophysical description of the electrical activity within muscle tissue and accounts for the coupling between the electrical potential in muscle fibers and the extracellular space. The model simulates the generation and propagation of MUAPs within the muscle tissue. MUAPs, which are the primary components of EMG signals, represent the electrical activity of individual MUs. By simulating MUAPs, the model can produce EMG signals that closely resemble those observed in experimental studies.

In the model, 15 baseline parameters specifying the physical and electrical properties of the muscle tissue are defined for the simulations, as summarized in Tab. III.1. The multi-domain model incorporates a cell membrane model by Shorten et al. (2007), with additional parameters and initial ion channel conductances, used to generate healthy virtual population summarized in Table 1 of the same paper. To simulate iEMG signals for each type of channelopathy, the chloride and sodium conductances ($g_{Cl}$ and $g_{Na}$, respectively) are adjusted to ensure that action potentials continue to be produced even after the electrical stimulus is removed, thus simulating myotonic discharges, as shown in Fig. III.3. The time-domain waveforms generated for both channelopathy types exhibit similar overall morphology.

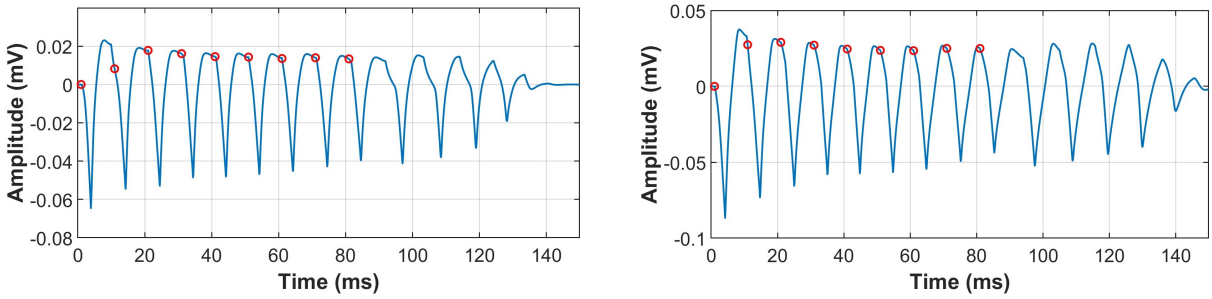
**Table III.1:** Baseline Parameters for the Channelopathy Simulation

| Parameter | Symbol | Value | Unit |
|---|---|---|---|
| Intracellular conductivity | $\sigma_i$ | $8.93 \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ | $\mathrm{mS\,cm^{-1}}$ |
| Extracellular conductivity (fiber direction) | $\sigma_e$ | 6.7 | $\mathrm{mS\,cm^{-1}}$ |
| Anisotropy of extracellular conductivity | $\sigma_e^{\text{aniso}}$ | 0.5 | – |
| Conductivity in fat | $\sigma_o$ | $0.4 \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ | $\mathrm{mS\,cm^{-1}}$ |
| Membrane capacitance | $C_m$ | 1 | $\mathrm{mS\,cm^{-2}}$ |
| Fiber surface-to-volume ratio | $A_m$ | 350 | – |
| Muscle length | $l$ | 4 | cm |
| Muscle width | $w$ | 1.5 | cm |
| Muscle height | $h$ | 1.5 | cm |
| Height of fat tissue | $h_{\text{fat}}$ | 0.5 | cm |
| MU territory center in y-direction | $y_{\text{center}}$ | 0.75 | cm |
| MU territory center in z-direction | $z_{\text{center}}$ | 0.75 | cm |
| MU territory radius[1] | $r$ | 0.2 | cm |
| MU territory load[2] | $load$ | 0.25 | – |
| Chloride channel conductance | $g_{\text{Cl}}$ | 4.5 | $\mathrm{mS\,cm^{-2}}$ |
| Sodium channel conductance | $g_{\text{Na}}$ | 0.0105 | $\mathrm{mS\,cm^{-2}}$ |

[1] MU territory radius represents the spatial extent of the MU territory within the muscle.
[2] MU territory load represents the fraction of the muscle volume that is occupied by the MU territory.

We created a population of 50 virtual subjects with parameters that vary around the baseline values for each of the groups representing healthy conditions, sodium and chloride channelopathy. Variations were introduced by multiplying the baseline parameters by noise factors following a normal distribution with a specified coefficient of variation (CoV). For the all of the parameters the CoV is 0.2, meaning the standard deviation of the noise is 20 % of the baseline value.

The simulated iEMG recordings were obtained at 3 different electrode positions, namely $(2.0, 0.5)$, $(2.5, 0.75)$, and $(3.0, 1.0)$ for each virtual subject. The total simulation time is 150 ms, with a time step of 0.01 ms. The output frequency reduces this by a factor of 10 by saving the results every 10th time step, resulting in 1500 time steps being recorded. The EMG potential is calculated at different depths along the z-axis for a given x and y electrode position. The depth is sampled at 17 points from the bottom to the top of the muscle grid, resulting in a matrix of size 17 by 1500 for each electrode position. This approach yields 51 samples per subject, which leads to a total of 2550 samples for each of the simulated channelopathy and healthy patient groups.




**(a)** Myotonic discharge (chloride-channel defect).

**(b)** Myotonic discharge (sodium-channel defect).

**Figure III.3:** Modeled myotonic discharges for the chloride-channel (left) and sodium-channel (right) defect classes. The red dots indicate stimulus application; the continued post-stimulus activity marks the myotonic discharge region. The healthy samples do not contain post-stimulus activity.

### III.1.3 Skeletal Muscle Fibrillation Potentials Dataset

The skeletal muscle fibrillation potentials data were collected at Eberhard Karls University of Tübingen between January 2021 and December 2023. The dataset includes 100 recordings of skeletal muscle fibrillations from 22 patients and 100 recordings of voluntary contractions from 19 patients. PD Dr. med. Justus Marquetand provided fully anonymized data for this study, approved by the ethics committee (project number: 056/2024BO1).

Fibrillation samples were recorded from the following muscles: right and left tibialis anterior, right biceps brachii, right deltoid, right vastus lateralis, right flexor digitorum profundus, left first dorsal interosseous, right and left abductor pollicis brevis, right extensor hallucis longus, left extensor digitorum longus, left extensor digitorum communis, left gastrocnemius, and left T10 paraspinal. Voluntary contraction recordings were obtained from the following muscles: right and left vastus lateralis, left gastrocnemius, right and left tibialis anterior, right pronator quadratus, left abductor digiti minimi, right first dorsal interosseous, right flexor pollicis longus, left S1 paraspinal, left biceps femoris, left infraspinatus, right and left deltoid, left serratus anterior, right extensor hallucis longus, left adductor longus, right biceps brachii, left vastus medialis, and left masseter.

The sampling frequency at which the data were collected is 64,000 Hz. The length of the recordings varies. Pre-filtering was performed using a high-pass filter with a cutoff frequency of 30 Hz and a low-pass filter with a cutoff frequency of 10,000 Hz.

### III.1.4 Hypothesis for Skeletal Muscle Disorders Dataset

#### Channelopathy Dataset Classification

Myotonic discharges in skeletal muscle channelopathies, specifically in non-dystrophic myotonias, arise from defects in sodium or chloride channels. Previous clinical neurophysiology protocols, such as short and long exercise tests combined with muscle

cooling, can help distinguish different patterns of sarcolemmal excitability (Fournier et al., 2004; Fournier et al., 2006; Matthews et al., 2010; McManis et al., 1986). However, these protocols and the resulting EMG interpretations often rely heavily on manual inspection and do not always unequivocally discriminate between sodium and chloride channel defect-induced myotonias (Matthews et al., 2010; Trip et al., 2008; Drost et al., 2015). Furthermore, iEMG signatures, such as myotonic discharge intervals and amplitude changes (Drost et al., 2015), have shown promising but not universally adopted utility in differentiating chloride from sodium channelopathies. This suggests that additional, possibly unknown, signal characteristics might exist, which could be revealed by more advanced and automated methods of signal analysis.

It was shown that pathophysiological mechanisms in skeletal muscle channelopathies lead to characteristic alterations in the electrophysiological behavior of muscle fibers (Cannon, 2015). Specifically, abnormal sodium channel function can result in persistent inward currents or slowed inactivation (Lehmann-Horn et al., 1999; Mitrović et al., 1995), while defective chloride conductance contributes to impaired repolarization and can shift the voltage dependence of channel opening (Lehmann-Horn et al., 1999; Wagner et al., 1998). Patch-clamp experiments have demonstrated that sodium channel mutations may exhibit persistent inward currents, slowed decay, or frequent channel re-openings (leading to "faulty inactivation"), whereas chloride channel mutations can show decreased open probability in the physiological voltage range (Lehmann-Horn et al., 1999). Clinically, these ion channel abnormalities manifest as myotonic discharges, that is, as prolonged or repetitive firing in EMG recordings (Fuglsang-Frederiksen, 2006). In summary, according to the Figure II.1, myotonic discharges in sodium and chloride channelopathies arise from:

- **Persistent sodium channel opening or slowed inactivation** leading to repeated muscle fiber depolarization.
- **Reduced chloride conductance or shifted gating** delaying or preventing full repolarization.

Such abnormal activity can lead to distinct EMG waveforms with potentially unique time-amplitude and spectral characteristics for each channel defect type. Based on the molecular findings, we hypothesize that differences in gating defects, that is "faulty inactivation" vs. "reduced open probability" could produce sufficiently distinctive signatures in macro-scale – the iEMG domain, to allow automated classification. Keeping this in mind, we formulate the following set of hypotheses (H):

**H1: Distinct iEMG Features.**

*Myotonic discharges arising from sodium and chloride channelopathies exhibit distinguishable temporal and spectral features in iEMG measurements, possibly beyond the classical manual metrics observed in prior test protocols.*

Rationale: Patch-clamp studies reveal different gating defects in sodium vs. chloride

channels (Lehmann-Horn et al., 1999), which likely lead to distinct myotonic discharge shapes. However, clinical exercise protocols (Matthews et al., 2010; Fournier et al., 2006) only partially differentiate these subtypes, often depending on manual observations.

**H2: Need for Automated Analysis.**
*An automated algorithm incorporating advanced signal processing, such as spectral analysis and machine learning methods, can use prior unidentified or more subtle EMG features to discriminate between sodium and chloride channel defects.*
Rationale: Manual methods risk observer bias and might fail to detect subtle features in the myotonic discharge waveforms (Drost et al., 2015). Automated pipelines could reduce variability and discover additional discriminative metrics.

**H3: Clinical Utility and Novel Criteria.**
*If automated EMG-based classification reliably separates chloride and sodium channelopathies, then new diagnostic criteria can be proposed for routine clinical use, facilitating faster and more accurate genetic testing.*
Rationale: Ambiguous cases require screening of both CLCN1 and SCN4A due to overlapping electrophysiological patterns (Matthews et al., 2010; Trip et al., 2008). An automated approach with high specificity would help direct genetic analysis and reduce reliance on repeated or inconclusive tests.

**H4: Pathophysiological Insight.**
*A robust classification framework can enhance understanding of how altered sodium or chloride conductances modify myotonic discharge shapes and durations, offering deeper insight into channelopathy pathophysiology.*
Rationale: Detailed electrophysiological characterization may reveal how different ion channel defects produce unique discharge characteristics, thereby clarifying the mechanisms of myotonia (Cannon, 2015; Lehmann-Horn et al., 1999).

To evaluate these hypotheses, we will follow the plan below:

**Step 1: Feature Extraction and Discovery.**

- We will employ advanced signal processing methods to prepare the skeletal muscle channelopathy dataset described in Subsection III.1.1 for automated classification. In parallel, we will apply the same analysis to the simulated dataset described in Subsection III.1.2 as a controlled test of our methods. The synthetic data will potentially also help us explore explicit distinguishing characteristics under controlled conditions.

**Step 2: Machine Learning Classification.**

- We will develop supervised learning models to classify channelopathies using the extracted features.

- We will systematically evaluate performance and conduct comparative analyses to assess which models and features yield the most robust results.
- We will additionally perform uncertainty assessments to quantify the reliability of each model's predictions.

**Step 3: Pathophysiological Interpretation.**

- We will relate the discovered features back to molecular and physiological mechanisms of sodium and chloride channel dysfunction.
- We will highlight any new insights gained regarding ion channel behavior and how these findings might inform future diagnostic or therapeutic approaches.

By integrating iEMG measurements with advanced computational methods, we aim to establish a more precise means of differentiating sodium and chloride channelopathies. This includes refining or proposing new diagnostic thresholds, beyond previously reported metrics, to improve the recognition of ambiguous cases (Matthews et al., 2010; Trip et al., 2008). Ultimately, our approach promises novel physiological insights into how each channel defect specifically alters discharge patterns and muscle excitability, thereby guiding genetic testing and informing potential pharmacological interventions.

### Skeletal Muscle Fibrillation Potentials Dataset Classification

The fibrillation dataset described in Subsection III.1.3 serves as an additional validation to test the universal applicability of the classification algorithm developed for skeletal muscle channelopathies. Differentiating fibrillation potentials from voluntary contractions can be challenging, particularly for less experienced neurologists, yet it is clinically critical to recognize these spontaneous denervation markers accurately.

Fibrillation potentials are spontaneous di- or triphasic potentials with a positive initial deflection, lasting 2–3 ms, and typically having amplitudes of 100–200 $\mu$V (Buchthal et al., 1966; Fuglsang-Frederiksen, 2006). Together with positive sharp-waves, these potentials provide evidence of ongoing muscle fiber irritability serve as key indicators of underlying neuromuscular pathology. However, subtle morphological overlaps or atypical forms can make the distinction of fibrillation potentials from small, low-amplitude MU potentials under mild voluntary activity more difficult (Barkhaus et al., 2021). Such overlaps can lead even experienced clinicians to risk misinterpretation (Barkhaus et al., 2021). This underscores the need for automated, objective methods to improve diagnostic confidence and reduce observer bias. Using the fibrillation potentials dataset, we will test the following hypotheses:

**H1: Distinct iEMG Features.**
*Fibrillation potentials exhibit sufficiently distinct temporal and spectral characteristics compared to voluntary contraction potentials to be robustly recognized by the same automated machine learning framework used for channelopathy classification.*

Rationale: Fibrillation potentials differ from voluntary contractions by their spontaneous nature, short duration, and characteristic amplitude range.

**H2: Universal Applicability of the Automated Classification Pipeline.**
*By applying the previously developed classification pipeline to the fibrillation dataset, we expect comparable or better performance relative to its performance on the channelopathy dataset, thus demonstrating the potential for a universal EMG analysis tool for identification of diverse neuromuscular disorders.*
Rationale:

- If our pipeline can reliably distinguish fibrillation from voluntary contraction potentials without extensive modifications, this will support a broader claim that our signal processing and machine learning algorithm can form a universal diagnostic framework for EMG-based identification of multiple skeletal muscle pathologies.
- Novice neurologists can misinterpret fibrillation potentials. Even though many these signals follow "textbook" morphology, atypical or overlapping waveforms exist (Barkhaus et al., 2021), thus requiring a more systematic, data-driven approach for their recognition. An automated approach can reduce this observer bias by systematically evaluating temporal and spectral signal features.

In order to test the hypothesis, we will follow the following plan:

**Step 1: Data Preparation:**

- Preprocess the fibrillation dataset using the same pipeline as in the channelopathy data.
- Identify clear segments of fibrillation potentials versus segments containing voluntary contractions.

**Step 2: Classification Algorithm Application:**

- Extract domain-specific features to match the approach used for channelopathy classification.
- Apply model(s) developed for channelopathy type classification to the fibrillation dataset.

**Step 3: Performance Evaluation:**

- Evaluate the classifier's performance in distinguishing fibrillation potentials from voluntary contractions.
- Compare the results with the performance achieved on the channelopathy dataset, examining whether classification robustness extends to fibrillations.

**Step 4: Conclusion and Generalizability:**

- Interpret results in light of the pipeline's universality: does it capture the unique morphology and spectral features of fibrillations effectively?

If performance on this fibrillation dataset remains strong, it will illustrate that the machine learning framework is not restricted to distinguishing sodium vs. chloride channelopathies alone. Instead, it can adapt to a broader range of EMG abnormalities, supporting the notion of a *universal* EMG analysis tool for diagnosing diverse neuromuscular disorders.

## III.2 Obstructive Sleep Apnea Dataset

### III.2.1 Obstructive Sleep Apnea Dataset

Fully anonymized and de-identified data of patients with OSA who underwent MAS treatment was provided for this project by the laboratory of Prof. Lynne Bilston at Neuroscience Research Australia (NeuRA), Sydney, Australia. For comprehensive details regarding the data collection protocols, patient consent procedures, exclusion criteria, and patient demographics readers are referred to the methods section of Brown et al. (2021) and Jugé et al. (2022).

This OSA dataset consists of multi-channel, overnight PSG recordings stored in `.edf` format. Each patient underwent a baseline PSG and a post-treatment PSG with a MAS in situ. All PSG measurements were scored by trained sleep technicians. Untreated OSA severity was determined based on the AHI, with mild OSA defined as an AHI more than 10 and less than or equal to 15 events per hour (patients with AHI less than or equal to 10 events per hour were not included in the study), moderate OSA as more than 15 and less than or equal to 30 events per hour, and severe OSA as more than 30 events per hour (Jugé et al., 2022).

The MAS treatment response classification followed the criteria outlined in (Jugé et al., 2022). Responders were defined as those achieving an AHI less than or equal to 10 events per hour and a reduction of at least 50 % in AHI from baseline to final PSG, or an AHI less than or equal to 5 events per hour regardless of baseline reduction percentage. Partial responders achieved at least a 50 % reduction in AHI but ended with a final AHI more than 10 events per hour. Non-responders had less than a 50 % reduction in AHI. Both machine-generated and manually annotated (human) scoring results were available for each patient, stored in `.rml` files. For this analysis, human scoring results were used. Due to the small number of partial responders in the dataset, these individuals were included in the 'Responder' class for the purposes of this study.

We analyzed three distinct OSA datasets, which were obtained at different times and locations. The first dataset, which we will denote as *OSAMAS*, comprised data from 41 patients, with 18 responders and 23 non-responders. The second dataset, named as *CRC*, included data from 37 patients, with 15 responders and 22 non-responders. The

third dataset, *PhysMAS*, consisted of 24 patients, of whom 14 were responders and 10 were non-responders. It should be noted that these datasets are imbalanced, reflecting a natural distribution of treatment outcomes.

The channel content and sampling rates varied both across and within datasets, largely due to the different acquisition locations (hospitals in Sydney and NeuRA) and the distinct operators responsible for signal acquisition. Typically, each `.edf` file contains between 24 and 35 measurement and technical channels, though the exact number may vary depending on the signal types included. These channels commonly include EEG and EOG signals to monitor brain activity and eye movements respectively, along with chin and leg EMG channels to record muscle activity and ECG channels to monitor heart function. Respiratory measurements capture airflow using pressure transducers and thermistors, and also include channels with respiratory rate, thoracic and abdominal effort. Snoring is recorded via a microphone, while end-tidal $CO_2$ monitoring tracks $CO_2$ levels in exhaled air, and oxygen saturation and pulse channels measure blood oxygen levels and heart rate. Body position or movement measurements, which track changes in the patient's posture during sleep, are also present, along with additional device-specific technical channels.

For this study, the channel of interest is the patient flow channel, which measures respiratory airflow during sleep. These measurements are obtained using a nasal pressure transducer that detects changes in air pressure, typically placed at the nose via a cannula. This specific channel was chosen because it provides direct information on airflow patterns. Further details on the rationale behind this choice and the subsequent data preprocessing steps can be found in Section III.2.2. Throughout this section, any references to the dataset refer to baseline (pre-treatment) measurements from the patient flow channel.

The raw airflow signals and their corresponding timestamps were extracted from the PSG recordings and saved in `.csv` format for each patient. Because the data were collected at different times and locations, sampling rates varied among and within the three OSA datasets. In the OSAMAS dataset, flow measurements ranged from 10 Hz to 256 Hz, with the majority of files being collected at 10 Hz and few at 32, 100, 200, or 256 Hz. The CRC dataset has a fixed sampling rate of 100 Hz for all patients, as well as the PhysMAS dataset maintains a sampling rate of 250 Hz for all samples. As a result, file sizes and total recording lengths differ across the three datasets and, in the case of OSAMAS, also within the same dataset.

To facilitate automated analysis, the flow signal was synchronized with scoring annotations stored in `.rml` files. These annotations capture information on sleep stages, respiratory events, and neural events identified by sleep technicians. By merging the extracted flow channel data with the corresponding annotations, synchronized patient flow datasets

were generated in `.csv` format. Each `.csv` file contains columns for the timestamp, the measured flow at a corresponding sleep stage and respiratory and/or neural event.

Tab. III.2 illustrates the header and the first two rows of a sample `.csv` file. The first column lists the time in seconds, and the second column provides the flow value in liters per second. The third column specifies the sleep stage, with REM referring to rapid eye movement sleep and NonREM1, NonREM2, NonREM3 corresponding to stages one, two, and three of non-REM sleep. The fourth column indicates the event family, and the fifth column classifies the type of event, for instance, an arousal for neural events or a hypopnea, mixed, obstructive, or central apnea for respiratory events. The sixth column contains a Boolean flag showing whether the event starts at that particular timestamp, and the seventh column displays the total duration of the event in seconds.

**Table III.2:** Synchronized Flow Data Overview

| Time (s) | Flow (l/min) | Stage | Family | Type | Onset | Duration (s) |
|---|---|---|---|---|---|---|
| 15878.37 | 20.034 | NonREM2 | Respiratory | Hypopnea | True | 44.5 |
| 15878.38 | 19.911 | NonREM2 | Respiratory | Hypopnea | False | 44.5 |

From these synchronized `.csv` files, several subsets were created to isolate the flow signal for specific sleep stages or event types. Tab. III.3 summarizes the different `.csv` files prepared for further analysis. The first column lists the file names, the second column describes which features are included, and the third column indicates which features are excluded from those files.

**Table III.3:** Overview of Flow Data Files

| File Name | Included Features | Excluded Features |
|---|---|---|
| `flow_data.csv` | Full patient flow | None |
| `flow_data_sleep_stages.csv` | Patient flow recorded during stages 1, 2, 3, and REM | Wake |
| `flow_data_REM.csv` | Patient flow recorded only during REM | Wake, Stages 1, 2, 3 |
| `flow_data_respiratory.csv` | Patient flow recorded only for respiratory events | All other events |

This structure allows to focus on specific aspects of the data, for example, to analyze only respiratory events or flow patterns during sleep. By structuring the data in this manner, we facilitate targeted investigations into how respiratory airflow patterns evolve under different conditions.

### III.2.2 Hypothesis for Obstructive Sleep Apnea Dataset

Our central hypothesis is that baseline respiratory airflow signals contain features predictive of patient response to MAS. By analyzing both time-domain and frequency-

domain characteristics of the flow signal, along with relevant PSG scoring information, we propose to develop an automated classifier capable of distinguishing potential responder patients from non-responders prior treatment. We will use the AHI reduction as our primary metric of treatment success, according to criteria described in section III.2.1.

MAS functions primarily by mechanically advancing the mandible to enlarge the upper airway, thereby reducing airway collapsibility. This approach often leads to fewer obstructive respiratory events and lower post-treatment AHI. Nevertheless, the degree of improvement varies substantially across individuals, likely due to anatomical differences that determine how well the mandible's forward repositioning stabilizes the velopharyngeal region (Chan et al., 2020).

Although we continue to employ AHI as our clinical endpoint, recent work by Mann et al. (2019) underscores how the *shape* of the airflow signal can reflect subtle but important anatomical and mechanical aspects of airway patency. Their findings suggest that continuous, shape-based metrics of flow provide a more nuanced view of obstruction severity than event-based indices alone. While we do not adopt their specific flow-ratio metric, the principle that flow-waveform morphology encodes information about airway mechanics motivates our strategy to exploit these shape features for classification.

We hypothesize that certain baseline flow patterns indicate an airway anatomy amenable to the mechanical expansion provided by MAS, manifesting in detectable differences in the flow signal prior to therapy. In normal breathing, low-frequency components (below 1 Hz) dominate, and the flow curve exhibits smooth inspiratory/expiratory phases. In OSA, partial collapse introduces inspiratory flow limitations, often seen as flattened or truncated inspiratory peaks and a shift of spectral energy to lower frequencies. Complete obstruction appears as near-total cessation of airflow (apneas), reflected by a loss of spectral power at the fundamental breathing frequency. The presence of high-frequency components, for example 20–100 Hz, in respiratory airflow signals, associated with snoring or recovery breaths following hypoventilation periods, may also mark dynamic airway narrowing.

Because MAS aims to mechanically modify such anatomical causes of obstruction, we expect that individuals whose baseline flow signals reveal prominent inspiratory flow restrictions (but are nonetheless susceptible to jaw advancement) will be more likely to demonstrate a significant drop in AHI after treatment. Consequently, flow-shape indicators and spectral features should serve as strong predictors of MAS response.

To test our hypothesis, we will:

1. **Acquire Baseline Respiratory Airflow Recordings:** We will construct multiple datasets from patient airflow data. Specifically, we will create:
    - **Sleep Data:** We will first extract the relevant part of the patient flow meas-

urements, that is data points corresponding to sleep that also includes periods with arousal.

- **All Respiratory Events:** A comprehensive dataset capturing all abnormal respiratory events during sleep: including apneas, hypopneas, and related arousals.
- **REM Stage:** We will extract airflow signals recorded during REM sleep. REM sleep is characterized by muscle atonia and irregular breathing patterns, and in patients with OSA, it has been consistently linked to more severe airway obstruction than non-REM sleep (Dempsey et al., 2010).

2. **Extract Time-Domain and Spectral Features:** We will analyze the acquired airflow signals to identify and quantify relevant time-domain amplitude features and spectral characteristics. These features are hypothesized to serve as predictive biomarkers of MAS treatment success.

3. **Develop a Machine-Learning Classifier:** Utilizing the extracted features, we will train a supervised machine-learning classifier to predict significant reductions in the AHI after MAS therapy. Post-treatment AHI scores will serve as the target labels for each data sample, thereby enabling the model to distinguish potential responders from non-responders.

Since the primary objective is to avoid withholding an effective treatment from those who could benefit, we will prioritize *sensitivity* (recall) as our key performance metric. By maximizing the correct identification of true responders, we minimize the risk of undertreatment, even if it leads to a higher false-positive rate.

In essence, we posit that the anatomy-driven effectiveness of MAS is reflected in the patient's baseline airflow morphology. By analyzing patient airflow data, alongside with conventional PSG indices, we anticipate identifying predictors of treatment response and building a classification model that accurately predicts MAS responders prior to treatment. Such a predictive model would allow clinicians to more confidently recommend MAS to those most likely to experience improvements in AHI.

## Chapter IV

# Signal Processing Methods and Analysis Techniques

The signal processing component of this project is dedicated to extracting the salient features embedded within biomedical signals – features that are critical for the subsequent classification tasks described in Chapter V. In this chapter, we establish both the theoretical foundations and the practical implementation details of the methods used to process our two distinct datasets: skeletal muscle iEMG signals and airflow recordings, extracted from polysomnography, described in Chapter III.

Although advanced neural network architectures such as Recurrent Neural Networks and Long Short-Term Memory networks can process raw signals directly, they typically require extensive computational resources and large volumes of training data. In contrast, our datasets are relatively small compared to the complexity and richness of the signal features. Consequently, we adopt a feature-engineering approach that leverages both time-domain (amplitude-based) and spectral (frequency-based) analyses to extract the most informative characteristics from the signals.

Furthermore, though deep neural networks can serve as automatic feature extractors, we have found it advantageous to pre-extract certain handcrafted features. This hybrid strategy not only reduces the computational burden on data-intensive neural architectures but also allows us to use more conventional machine learning algorithms, ultimately maximizing both interpretability and classification performance.

This chapter is structured as follows:

- **Part I (§IV.1–IV.2):** We begin with a discussion of signal sampling and digitization, reviewing key concepts such as sampling frequency, aliasing, and the properties of

linear time-invariant systems. Next, we explore fundamental time-domain analyses, including statistical metrics used to characterize amplitude variability and periodicity, and then address noise reduction techniques, emphasizing the role, design trade-offs, and limitations of digital filtering.

- **Part II (§IV.3):** We extend the discussion to frequency-domain methods by first defining the signal spectrum and the short-time Fourier transform, and then presenting advanced approaches, such as the continuous wavelet transform and superlet transform, that offer adaptive resolution for non-stationary signals. We also explain how one-dimensional signals can be transformed into two-dimensional scalograms suitable for image-based feature extraction.

- **Part III (§IV.4):** We present image processing methodologies, including texture analysis and the gray-level co-occurrence matrix.

- **Part IV (§IV.5–IV.6):** We apply the aforementioned techniques to our datasets and thus demonstrate feature extraction procedures and obtain both conventional time-domain and more advanced image-based representations of the signals.

# IV.1 Fundamentals of Biomedical Signal Processing

## IV.1.1 Discrete-Time Signals and Sampling Theorem

A signal is a function that conveys information about a phenomenon. In many practical applications of biomedical engineering, signals are not handled in continuous form but rather as *discrete-time signals*, sampled at regular intervals. The duration of these intervals is the *sampling period* $T_s$ (seconds per sample), and its reciprocal $f_s = 1/T_s$ (samples per second, Hz) denotes the *sampling frequency* or in other words, *sampling rate* of a signal. If the discrete-time samples are indexed by an integer $n$, then a discrete-time signal can be expressed as $x[n]$.

Shannon's sampling theorem states that, to prevent aliasing – where higher-frequency components of a band-limited signal appear artificially at lower frequencies in the sampled data – one must sample the signal at a rate at least twice the signal's highest frequency component (Shannon, 1948). Formally, if $f_m$ is the maximum frequency present in the analog signal, then the sampling rate must satisfy $f_s \geq 2f_m$. When the signal is not naturally band-limited, a low-pass filter is typically applied in hardware to remove any frequency components above half the sampling rate – known as the Nyquist rate ($\frac{f_s}{2}$) – thereby preventing aliasing artifacts (Cerutti et al., 2011).

Sampling rate choices critically affect which features can be reliably extracted from biomedical data. For example, EMG recordings of muscle activity may require sampling rate of 1–5 kHz or higher to capture the rapid action potentials without aliasing. Our muscle channelopathy iEMG data was acquired at 44.1 kHz, ensuring faithful

representation of high-frequency myotonic discharges. Conversely, OSA airflow signals, sampled between 10 Hz and 250 Hz, may fail to capture abrupt events or snoring components if their fundamental frequencies exceed the available sensor bandwidth. The sampling frequency heavily influences which signal characteristics can be reliably measured. Throughout this thesis, we will observe how the chosen sampling frequency shapes the feature extraction process and, ultimately, influences the effectiveness of classification tasks.

### IV.1.2 Linear Time-Invariant Systems

Digital signal processing (DSP) typically operates on discrete-time, discrete-amplitude signals, where each sample is acquired at a fixed rate and subsequently *quantized*, that is rounded to match the limited numerical precision available in the machine (Cerutti et al., 2011). In practice, these discrete signals are passed through a *linear time-invariant* (LTI) system: a *system*, in this context, refers to any process that produces an output upon receiving an input (Smith, 1997).

A system is called *linear* if it satisfies two basic properties: homogeneity and additivity. Homogeneity means that scaling the input by a constant $k$ scales the output by the same constant $k$. Additivity implies that when multiple inputs are summed, the output is simply the sum of the individual outputs corresponding to each input. In addition, a system is said to be *time-invariant* if a shift in the input signal results in an identical shift in the output, which indicates that the system's behavior remains unchanged over time (Smith, 1997).

An LTI system is fully described by its *impulse response*, $h[n]$. When the input to the system is a *discrete delta function*, $\delta[n]$ – which is zero for all $n \neq 0$ and equals 1 at $n = 0$ – the output directly results in the impulse response. More generally, any input $x[n]$ can be expressed as a combination of shifted and scaled delta functions, each of which produces a correspondingly shifted and scaled version of the impulse response. Therefore, the output for a general input $x[n]$ is given by the convolution of $x[n]$ with the system's impulse response $h[n]$, denoted as $x[n] * h[n]$. Let the length of $h[n]$ be $K$. The output $y[n]$ in response to $x[n]$ is then given by

$$y[n] \;=\; x[n] * h[n] \;=\; \sum_{k=0}^{K-1} h[k]\, x[n-k], \tag{IV.1}$$

where $h[k]$ quantifies how the system responds to an impulse at time index zero. By combining shifted and scaled versions of this impulse response, the system effectively processes more general input signals.

Within DSP terminology, the impulse response $h[k]$ is often referred to as a *digital filter kernel* (Smith, 1997). By appropriately designing $h[k]$, it becomes possible to

attenuate certain signal components, such as noise or baseline drift, while retaining or enhancing features of interest. This fundamental viewpoint guides many advanced techniques described later in this chapter, including filtering and other transformations in Sections IV.2 and IV.3.

### IV.1.3 Time-Domain Characteristics of Signal

In this section, we focus on *time-domain* characteristics, which describe how a signal evolves over time. In discrete-time form, some signals are effectively nonzero only within a finite range of indices. Let $x[n]$ be defined for integer $n$. If $x[n] \neq 0$ only for $n_1 \leq n \leq n_2$, then its *support interval* is $[n_1, n_2]$. The length of this support interval, $(n_2 - n_1 + 1)$, is called the signal's *duration*. In practice, many measured signals are assumed to be zero outside this range for computational convenience.

Another fundamental signal property is *periodicity*. A discrete-time signal $x[n]$ is said to be periodic with period $N$ if $x[n + N] = x[n]$ for all $n$. The smallest positive integer $N$ that satisfies this relationship is called the *fundamental period*. However, many biomedical signals often display transient or quasi-periodic patterns. For example, an EMG recording may show bursts of repetitive firing, yet the exact shape or timing of each burst can vary significantly. In other cases, signals may be *aperiodic*, that is extending indefinitely over the discrete index range from $-\infty$ to $+\infty$ without repeating in a regular pattern. Formally, such a signal does not satisfy $x[n + N] = x[n]$ for any positive integer $N$ and all integer $n$.

Beyond these basic descriptors, a variety of *time-domain signal statistics* exist to provide insight into the signal's amplitude distribution, variability, or oscillatory nature. We will list some of the such common measures.

*Average (mean) value* represents the baseline level of a signal and is defined as

$$\mu = \frac{1}{N} \sum_{n=0}^{N-1} x[n], \tag{IV.2}$$

For an EMG recording, for example, a shift in the mean value may indicate unusual baseline muscle activity or an overall increase in tone.

*Variance* (VAR) captures how spread out the signal values are around the mean:

$$\sigma^2 = \frac{1}{N} \sum_{n=0}^{N-1} \big(x[n] - \mu\big)^2, \tag{IV.3}$$

A higher variance in an EMG signal can reflect greater MU recruitment or more volatile

muscle activation. The *standard deviation* (std), $\sigma$, is the square root of the variance and thus shares the same units as the original signal, making it a more intuitive measure of the overall fluctuations.

*Zero Crossings* (ZC) provides a simple yet effective frequency-oriented descriptor by counting how often the signal crosses the zero-amplitude axis (Hudgins et al., 1993). To mitigate noise-induced sign changes, a small threshold $\Delta$ is introduced, yielding:

$$\mathrm{ZC} = \sum_{k=1}^{N-1} \mathbf{1}([x_k x_{k+1} < 0] \wedge |x_k - x_{k+1}| \geq \Delta)\,, \tag{IV.4}$$

where $x_k$ and $x_{k+1}$ are consecutive samples, and $\mathbf{1}\{\cdot\}$ is an indicator function. By ignoring sign inversions below the threshold $\Delta$, this approach reduces spurious zero crossings caused by baseline drift or low-amplitude noise. In practice, a higher ZC reflects more rapid oscillations, potentially indicating bursts of activity or abrupt waveform changes in signals.

*Energy* quantifies the total "strength" of a signal over its duration:

$$E = \sum_{n=-\infty}^{\infty} |x[n]|^2, \tag{IV.5}$$

When analyzing signals that do not decay over time like, for example, infinite sinusoids, the total energy $E$ becomes infinite, making these signals *not squarely summable*. In such cases, the *average power* serves as a more meaningful characteristic. For a discrete-time signal $x[n]$, the average power $P_x$ is defined as:

$$P_x = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \big|x[n]\big|^2. \tag{IV.6}$$

Based on these definitions, signals can be classified as either *finite-energy signals* (where $0 < E < \infty$ and total power is zero) or *finite-power signals* (where $0 < P < \infty$ and energy is infinite).

In certain situations, more sophisticated statistical metrics can offer deeper insights into a signal's distribution and its sensitivity to outliers. Three such higher-order features commonly considered in signal analysis are *shape factor*, *kurtosis*, and *skewness*.

*Shape factor* describes how spread out a waveform is, independently of amplitude scaling.

To define shape factor, we first compute the *root-mean-square* (RMS) value of the signal:

$$x_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\big(x[n]\big)^2}. \tag{IV.7}$$

The shape factor is then the ratio of the RMS and the mean absolute value of the signal:

$$\text{shape factor} = \frac{x_{\text{rms}}}{\frac{1}{N}\sum_{n=1}^{N}|x[n]|}. \tag{IV.8}$$

If a signal has sharper peaks, its shape factor will be larger, whereas flatter or more uniformly distributed waveforms result in lower shape factor values.

*Kurtosis* is a statistical measure reflecting the tendency of the signal's amplitude distribution to produce outliers (Krishnan, 2021). Specifically, kurtosis assesses the extremity of deviations from the mean, particularly higher kurtosis implies that large-amplitude events occur more frequently than they would in a normal distribution. For a discrete-time signal $x[n]$ of length $N$ with mean $\overline{x}$,

$$\text{kurtosis}(x) = \frac{\frac{1}{N}\sum_{n=1}^{N}\big(x[n]-\overline{x}\big)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}\big(x[n]-\overline{x}\big)^2\right)^2}. \tag{IV.9}$$

A normal distribution has a kurtosis of 3. Values exceeding 3 indicate heavier tails in the distribution and hence more frequent extreme values.

*Skewness* quantifies the asymmetry of a signal's amplitude distribution and can help detect the presence of outliers. A signal with positive skewness has a distribution skewed toward higher amplitudes, indicating a longer or heavier tail on the right; conversely, negative skewness suggests a heavier tail on the left. A skewness value of zero signifies a perfectly symmetrical distribution. The skewness for a discrete-time signal $x[n]$ of length $N$ is:

$$\text{skewness}(x) = \frac{\frac{1}{N}\sum_{n=1}^{N}\big(x[n]-\overline{x}\big)^3}{\left(\frac{1}{N}\sum_{n=1}^{N}\big(x[n]-\overline{x}\big)^2\right)^{3/2}}. \tag{IV.10}$$

Such higher-order statistics are valuable for identifying abnormal signal patterns or outliers in the amplitude distribution, as they highlight deviations from normal behavior.

Beyond simple statistical metrics, *time-domain* analysis often includes *autocorrelation*, which compares a signal with a time-shifted version of itself (Krishnan, 2021). Autocorrelation can be viewed as the convolution of a signal $x[n]$ with a reversed copy of itself, making it well-suited for detecting periodic or repetitive structures. In discrete-time form,

one common definition of the autocorrelation function is:

$$R_x[k] = \sum_{n=0}^{N-1-k} x[n]\, x[n+k], \qquad k \geq 0, \tag{IV.11}$$

where $k$ indicates the time shift. This metric is especially useful for identifying periodic or repetitive structures within a signal, such as repetitive firing activity in EMG recordings.

While time-domain features can reveal certain signal characteristics, relying solely on them for highly non-stationary signals with diverse frequency content often limits discriminative power in complex classification tasks. Consequently, advanced methods such as spectral analysis are typically used to complement time-domain approaches, as they better account for the dynamic nature of biomedical signals. In the following sections, we elaborate on these concepts by introducing time-frequency representations and additional feature extraction techniques.

## IV.2 Noise Reduction and Digital Filtering

While time-domain analyses offer an initial view of how biomedical signals evolve over discrete time samples, real-world measurements are often contaminated by noise or interference, obscuring meaningful information. Digital filtering provides a direct way to mitigate such contamination, preventing it from masking physiologically important information. Building on the notion of LTI systems introduced in Section IV.1.2, this section illustrates how convolution with suitably designed filter kernels can suppress unwanted signal components, preserve features of interest, and pave the way toward more advanced frequency-based methods.

Filters are typically designed to accomplish two main objectives. One is *signal separation*, meaning the removal of interference or extraction of desired components, while the other is *signal restoration*, in which the filter corrects distortions in the recorded data (Smith, 1997). As an example, a high-pass filter can eliminate low-frequency movement artifacts in EMG recordings. We begin with a discussion of signal-to-noise ratio (SNR) as a metric for evaluating how effectively a filter improves data clarity, and then explore the time- and frequency-domain representations of filters, along with the crucial design compromises inherent in filtering.

### IV.2.1 Signal-to-Noise Ratio

The *SNR* provides a quantitative measure of how distinguishable the desired signal is from background noise. If $P_{\text{signal}}$ and $P_{\text{noise}}$ represent the respective powers of the signal and noise, the SNR (in decibels) is defined by

$$\text{SNR (dB)} = 10 \log_{10}\Big(\frac{P_{\text{signal}}}{P_{\text{noise}}}\Big). \tag{IV.12}$$

A higher SNR indicates a clearer separation between signal and noise, making it easier to isolate and interpret relevant patterns in a signal. Monitoring how the SNR changes under various filtering configurations helps determine whether a chosen filter truly enhances data quality or inadvertently removes important features.

### IV.2.2 Digital Filters

Digital filters are commonly modeled as LTI systems, producing an output $y[n]$ in response to an input $x[n]$ through convolution with an impulse response $h[n]$, as shown in Equation IV.1 of the Section IV.1. When $h[n]$ is nonzero only over a finite range, the filter is called a finite impulse response (FIR) filter. FIR filters are unconditionally stable and can maintain linear phase, meaning they shift and scale time-domain waveforms with minimal distortion. However, achieving steep transitions in the frequency domain often necessitates a longer impulse response, leading to higher computational costs. In contrast, infinite impulse response (IIR) filters rely partly on previously computed outputs to generate the current output. This approach typically requires fewer coefficients to attain similar levels of frequency selectivity but may introduce non-linear phase behavior or demand extra care to ensure stability (Smith, 1997; Oppenheim, 1999).

Although impulse response reveals how a filter responds to a single impulse, every linear filter also has a step response and a frequency response, each providing an alternative perspective on the filter's behavior. Specifying any one of these three representations fixes the other two, because they describe the same system in different ways. Taken together, they show how the filter behaves across various conditions, making all three indispensable for understanding filter operation (Smith, 1997).

The *step response* provides a time-domain perspective on how a filter behaves when the input signal undergoes an abrupt transition from zero to a constant amplitude, in other words jumps from zero to a fixed value. Observing key features such as the rise time, which indicates how quickly the output transitions from zero to its final level, overshoot, which reveals whether the output briefly exceeds that final level, and the symmetry of the response, can offer valuable insight into a filter's ability to handle transient phenomena without significant distortion. Step decomposition, which divides an $N$-sample input into $N$ constant-amplitude segments added at specific time indices (Smith, 1997), further illustrates this concept by showing how each "step" component passes through the filter. If, for instance, certain frequencies or amplitudes tend to be amplified or attenuated, it becomes easier to identify systematic distortions in the time-domain waveform.

While the step response focuses on time-domain performance, the *frequency response*, denoted $H(\omega)$, reveals how strongly different frequency components of the input are passed, attenuated, or shifted in phase (Smith, 1997). In essence, convolving the input $x[n]$ with the impulse response $h[n]$ in the time domain corresponds to multiplying the Fourier

transform $X(\omega)$ by $H(\omega)$ in the frequency domain to produce $Y(\omega)$, as summarized by:

$$x[n] * h[n] \quad \Longleftrightarrow \quad X(\omega)\, H(\omega). \tag{IV.13}$$

In discrete-time systems, this frequency response can be derived by evaluating the filter's transfer function $H(z)$ on the unit circle ($z = e^{j\omega T}$). Equivalently, the frequency response can be viewed as the Discrete-Time Fourier Transform of $h[n]$:

$$H\big(e^{j\omega T_s}\big) = \sum_{n=-\infty}^{\infty} h[n]\, e^{-j\omega n T_s}, \tag{IV.14}$$

which reveals how each sinusoidal component of frequency $\omega$ is scaled and phase-shifted by the filter. Specific details on how this transform connects with spectrum estimation and broader spectral analysis concepts appear in Section IV.3, where the role of frequency-domain methods in analysis is explored more extensively.

### IV.2.3 Filtering Approaches in Digital Signal Processing

Digital filters are often classified according to the frequency regions they target in a given signal. A low-pass filter removes high-frequency components or rapid oscillations, which is useful for smoothing signals and eliminating high-frequency noise. Conversely, a high-pass filter preserves faster dynamics by blocking low-frequency drifts or baseline wander. In some cases, band-pass filters are used to retain only the activity within a specific frequency band, for instance isolating the 20–500 Hz range in EMG data where muscle activity predominantly appears. Filters can also be designed to reject specific narrow spectral bands, known as band-reject or notch filters, which are useful for eliminating disturbances such as the 60 Hz power line interference (Nilsson et al., 1993) commonly encountered in biomedical recordings.

From a time-domain perspective, implementing a filter means convolving the input $x[n]$ with a kernel $h[n]$. In the frequency domain, the same operation appears as multiplying the signal spectrum $X(\omega)$ by the filter response $H(\omega)$. Both viewpoints guide the filter's design (Smith, 1997). One might begin by specifying a desired passband and stopband in the frequency domain, then examine how the resulting filter behaves in the time domain. For example, a gentler cutoff transition can mitigate overshoot and ringing but may allow some undesired frequencies to leak through, whereas a sharper transition band typically requires a longer impulse response, leading to higher computational demands and potentially more pronounced time-domain ringing that distorts brief events.

These trade-offs illustrate why time and frequency perspectives are complementary in filter design. Time-domain criteria focus on how the filter responds to sudden changes, such as whether it accentuates or suppresses rapid events, and whether overshoot or ringing appears. Frequency-domain criteria specify how strongly certain components must be

attenuated or passed, as well as the extent of phase distortion that may shift waveforms in time or alter their shape. Designers generally iterate between these domains, adjusting a filter's frequency-domain specifications and evaluating the resulting time-domain behavior until a workable balance is struck. Although the ideal "brick-wall" filter, featuring an abrupt rectangular cutoff, appears attractive in the frequency domain, it cannot be realized in practice without an infinitely long impulse response (Smith, 1997). As a result, filter design inevitably involves trading off sharpness, computational efficiency, and artifact suppression to preserve the essential features of a signal while minimizing unwanted distortions.

### IV.2.4 Advantages and Disadvantages of Filtering

A well-chosen filter can substantially improve a signal's clarity when the noise frequencies are sufficiently distinct from the signal's primary band (Smith, 1997). However, if the signal frequencies overlap with those of the noise, filtering may also erase valuable features along with undesired components. As previuously discussed, overly aggressive cutoff transitions can distort rapid transients or induce ringing, while too lenient an approach might allow excessive noise.

In many practical measurement setups, such as EMG or ECG acquisitions, analog filters built into the hardware amplifier often remove extreme out-of-band noise or shift the baseline before digitization. Digital filtering can then refine the signal further, provided that the relevant frequency content is well-defined. However, any filter choice must be approached carefully. Narrow filtering bands risk excluding signal details, whereas insufficient filtering may hinder robust event detection.

When classifying complex phenomena, tracking changes in the SNR under various filter settings can help confirm whether filtering truly enhances the features of interest or inadvertently suppresses them. In practice, an excessively narrow passband may conceal meaningful information, while minimal filtering might leave the data susceptible to noise. Our iEMG data, illustrate this balancing act: hardware amplifiers already imposed certain limits, and we further applied software-based band-pass filtering, as it will be discussed in the Section IV.5. Generally, we tried to keep additional filtering minimal to avoid losing potentially relevant high-frequency cues, letting classification algorithms identify the most salient bands.

Digital filtering can indeed be highly effective in reducing interference and shaping signals in ways that preserve fundamental biomedical patterns. However, filters are invariably fixed in their spectral characteristics, making them less appropriate if a signal's main frequencies shift substantially over time. In the following sections, we expand these ideas into broader frequency-domain analysis and time-frequency methods, which allow a more

adaptive and detailed exploration of non-stationary biomedical signals.

## IV.3 Frequency-Domain and Spectral Analysis

The *frequency domain* of the signal provides a complementary perspective to its *time domain*. Whereas the time-domain representation focuses on how a signal evolves over time, the frequency domain shows how its energy or power is distributed across different frequencies. Because physiological processes often occupy characteristic frequency ranges, frequency-domain analysis is an invaluable method for revealing hidden patterns, discarding irrelevant content, and setting the stage for more advanced approaches when signals are highly non-stationary.

### IV.3.1 Spectrum: Definition and Importance

The term *spectrum* refers to the frequency content of a signal. When a discrete-time signal $x[n]$ is decomposed into its sinusoidal components, each component occupies a distinct frequency band. The recombination or *Fourier synthesis* of these components precisely reconstructs the original signal (Proakis, 2001). Conversely, omitting or altering any frequency component produces a different signal. This indicates that the spectrum effectively acts as the signal's "signature".

Converting a signal from the time domain to the frequency domain is termed *analysis*, whereas reconstructing the time-domain signal from these frequency components is referred to as *synthesis*. In discrete-time settings, these operations often rely on two closely related transforms: the *Discrete-Time Fourier Transform* (DTFT) and the *Discrete Fourier Transform* (DFT). The DTFT is applied to aperiodic signals, resulting in a function $X\left(e^{j\omega}\right)$ that is always $2\pi$-periodic with respect to the continuous frequency variable $\omega$. Formally, the DTFT of $x[n]$ is defined by

$$X\left(e^{j\omega}\right) \;=\; \sum_{n=-\infty}^{\infty} x[n]\, e^{-j\,\omega\,n}, \quad -\pi \le \omega < \pi, \tag{IV.15}$$

where $\omega$ is the angular frequency in radians/sample (Ling, 2010).

Because an infinite number of sinusoids is required to perfectly reconstruct an aperiodic sequence, it is generally impractical to compute the DTFT in most computer algorithms (Smith, 1997). Instead the *Discrete Fourier Transform* (DFT), which applies to periodic discrete-time signals, is used for practical purposes. In this case, one treats $x[n]$ as either one period of an $N$-periodic sequence or simply a finite record of data from $n = 0$ to $n = N - 1$. The DFT then maps the time-domain sequence of length $N$ to a set of

frequency-domain samples:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, \ldots, N-1, \tag{IV.16}$$

where $\frac{2\pi}{N}$ is the fundamental frequency resolution in radians/sample for the DFT, and $k$ indexes the discrete frequency bins (Ling, 2010). Each output $X[k]$ is a complex-valued coefficient associated with the discrete frequency $\omega_k = \frac{2\pi}{N}k$. The inverse DFT reconstructs the original sequence from these frequency coefficients.

From an computational viewpoint, determining the frequency content mathematically is called *spectral analysis*. However, in practical measurement settings, for example for capturing EMG data from sensors, one relies on *spectrum estimation* to approximate the signal's frequency content from sampled data (Proakis, 2001). Whether analytical or empirical, identifying dominant frequency components is often crucial for revealing physiologically relevant features of the signal.

One of the simplest approaches to estimate the power spectrum of a discrete-time signal is constructing a *periodogram*. Given a finite window of signal samples, the periodogram is obtained by computing the DFT of the windowed data, taking the squared magnitude of the result, and then normalizing by the segment length $N$:

$$\text{Periodogram}(f_n) \;=\; \frac{1}{N}\big|\text{DFT}\{x[n]\}(f_n)\big|^2, \tag{IV.17}$$

where $f_n$ denotes discrete frequency samples. Although straightforward, a single periodogram can exhibit high variance and is sensitive to noise, especially if the signal is short or contains non-stationary behavior.

### Welch's Method for Power Spectrum Estimation

An alternative or supplementary approach to single-periodogram methods is *Welch's method* (Welch, 1967), which reduces variance in power spectrum estimates by segmenting and averaging multiple periodograms. This nonparametric technique for power spectral density (PSD) estimation produces a smoother, lower-variance power spectrum. A PSD plot typically shows estimated signal power or energy per unit time/frequency on the vertical axis against frequency on the horizontal axis. Such a visualization enables identification of how signal power is distributed over different frequency bands.

In Welch's method, the signal is divided into $K$ overlapping segments of length $L$, commonly with 50% overlap between adjacent segments. Each segment is then multiplied by a window function $W(j)$ to mitigate edge effects. The DFT of each windowed segment provides the

frequency samples $A_k(n)$. The resulting periodogram for the $k$th segment is

$$I_k(f_n) \;=\; \frac{1}{L\,U}\big|A_k(n)\big|^2, \tag{IV.18}$$

where $U$ is a normalization constant, which is often the sum of squared window coefficients. Averaging the $K$ periodograms produces Welch's PSD estimate:

$$P_{\text{Welch}}(f_n) \;=\; \frac{1}{K}\sum_{k=1}^{K} I_k(f_n). \tag{IV.19}$$

Averaging multiple segments significantly reduces the variance of the PSD estimate compared to a single periodogram computed from the entire record (Welch, 1967). This procedure also offers flexibility in selecting segment lengths and overlaps, thus balancing spectral resolution, variance reduction, and computational efficiency.

In our analyses of both channelopathy and OSA datasets, we employed Welch's method to obtain power spectral estimates. For the EMG dataset, PSD plots were used to identify relevant muscle activity frequency bands, particularly for signal filtering purposes. In the OSA dataset, we compared Welch PSDs of responders and non-responders to the treatment with MAS to examine spectral differences in respiratory airflow signals across classes. These PSD-based features allowed us to quantify the dominant frequencies and overall power distributions in each class.

**Energy Equivalence – Parseval's Relation.**

A fundamental property of frequency-domain analysis is that it preserves the total signal energy when transforming between time and frequency domains. In the context of the DTFT, Parseval's relation can be written as

$$\sum_{n=-\infty}^{\infty}\big|x[n]\big|^2 \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big|X\big(e^{j\omega}\big)\Big|^2\,d\omega. \tag{IV.20}$$

where $x[n]$ is a discrete-time signal and $X(e^{j\omega})$ is its DTFT. Intuitively, this identity means that transforming a signal into the frequency domain does not alter its total energy; one can integrate power either over $n$ (time) or $\omega$ (frequency) and arrive at the same result.

Suppose that discrete periodic signal, $x[n]$ is defined for $n = 0, \ldots, N-1$, and $X[k]$ is its $N$-point DFT. Then Parseval's relation in DFT analysis setting can be written as:

$$\sum_{n=0}^{N-1}|x[n]|^2 \;=\; \frac{1}{N}\sum_{k=0}^{N-1}|X[k]|^2, \tag{IV.21}$$

assuming the DFT's inverse transform has no additional normalization factors (Smith, 1997). If a different normalization is used for example, $1/\sqrt{N}$ in both the forward and inverse DFT, the equation would include corresponding scale factors on both sides. Nevertheless, the core principle remains: the total energy in the time domain matches the total energy computed in the frequency domain. In other words, Parseval's relation tells us that analyzing a signal's energy or power in the frequency domain is equivalent to analyzing it in the time domain.

In biomedical signal analysis, the primary objective is to isolate physiologically meaningful information from the various components that make up a recorded signal. For example, an EMG recording may be contaminated by 60 Hz power-line interference (Nilsson et al., 1993). By examining the signal in the frequency domain, this unwanted 60 Hz component can be distinguished from the muscle's main frequency band and selectively attenuated. Such clear separation of noise from physiological activity illustrates the importance of spectral analysis. Moreover, because specific physiological processes occupy particular frequency ranges, signals falling outside these ranges can often be removed without compromising clinically relevant information (Cerutti et al., 2011).

Nevertheless, many biomedical signals are *non-stationary*, meaning their frequency content varies over time. A purely frequency-based transform like the DFT can provide a *global* representation of the signal's spectrum but cannot indicate when those frequencies appear or shift within the recording. This limitation motivates *time-frequency* approaches, which capture both time and frequency information in a unified representation.

### IV.3.2 Short-Time Fourier Transform and Spectrograms

The *Short-Time Fourier Transform* (STFT) offers a practical means to handle signals whose frequency content changes over time (Cerutti et al., 2011). Rather than applying the Fourier transform over the entire record, and thus assuming stationarity, the STFT divides the signal into short time frames and computes a spectrum within each localized window. This approach is based on the fact that within small segments, the signal may be treated as *approximately* stationary, allowing a standard Fourier analysis:

$$\mathrm{STFT}_x(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]\, w[m-n]\, e^{-j\omega m}. \tag{IV.22}$$

where $w[m-n]$ is a window function, centered around time index $n$. By sliding this window across different $n$-values, one obtains a local spectrum for each time interval, thereby forming a sequence of frequency spectra that reveal how the signal's frequency content evolves over time.

### Spectrogram

As each STFT frame yields a *local* spectrum for a given time $n$, assembling these frames produces a *spectrogram*, which is a visual representation of the STFT, displaying the magnitude of the STFT coefficients as a function of time and frequency:

$$\text{Spectrogram}(n, \omega) \;=\; \left|\text{STFT}\{x[m]\}(n, \omega)\right|^2. \tag{IV.23}$$

This representation is typically visualized with time on the horizontal axis and frequency on the vertical axis, using color or intensity to reflect amplitude.

### Limitations of Short-Time Fourier Transform for Non-Stationary Signals

Although the STFT improves upon pure Fourier analysis by partitioning a signal into quasi-stationary segments, its time-frequency resolution remains tied to the chosen window length. Real-world biomedical measurements, such as those from muscle channelopathies or dynamic respiratory events, often exhibit rapid bursts of high-frequency activity together with slower oscillations, requiring more adaptable methods. A shorter window sharpens time resolution but coarsens frequency discrimination; a longer window refines frequency resolution at the expense of time localization. Because many signals evolve at multiple time and frequency scales simultaneously, no single window can optimally capture all relevant patterns, potentially obscuring critical features.

### Uncertainty Principle and Resolution Trade-Off

A rigorous way to see why the STFT cannot achieve arbitrarily fine resolution in both time and frequency is through *Heisenberg's uncertainty principle* (Slepian, 1983). In its simplest form for continuous-time analysis, this principle asserts that if $g(t)$ is a window function of finite energy, then its effective time and frequency spreads, denoted $\Delta t$ and $\Delta f$, cannot both be made arbitrarily small. Specifically,

$$\Delta t \, \Delta f \;\geq\; \frac{1}{4\pi}, \tag{IV.24}$$

where the equality holds only if $g(t)$ is a Gaussian (Champeney, 1987; Rioul et al., 1991).

To obtain (IV.24), one generally relies on a proof involving the Cauchy-Schwarz inequality, see (Percival et al., 1993), pages 72–75 or (Champeney, 1987), pages 75–77 for further details. In the standard derivation, $\Delta t$ and $\Delta f$ are defined as the RMS spreads of the window function $g(t)$ in time and of its Fourier transform $G(f)$ in frequency:

$$\Delta t \;=\; \sqrt{\frac{\displaystyle\int_{-\infty}^{\infty} t^2 \, |g(t)|^2 \, dt}{\displaystyle\int_{-\infty}^{\infty} |g(t)|^2 \, dt}}, \qquad \Delta f \;=\; \sqrt{\frac{\displaystyle\int_{-\infty}^{\infty} f^2 \, |G(f)|^2 \, df}{\displaystyle\int_{-\infty}^{\infty} |G(f)|^2 \, df}}, \tag{IV.25}$$

where $G(f)$ is the Fourier transform of $g(t)$ (Rioul et al., 1991).

The integrals in (IV.25) measure how the energy of $g(t)$ and $G(f)$ is distributed around their respective centers (often taken as the first moment, or "mean," in time and frequency). If $g(t)$ is tightly concentrated in time (making $\Delta t$ small), its transform $G(f)$ must inevitably spread out in frequency (making $\Delta f$ large), and vice versa. This trade-off is precisely why one cannot design a single window that is arbitrarily narrow in both time and frequency.

The equality in (IV.24) can only be attained by a Gaussian function. Intuitively, the Gaussian's symmetrical shape and rapid exponential decay balance its energy "evenly" between time and frequency domains, minimizing the product $\Delta t\,\Delta f$. This makes Gaussians theoretically optimal for STFT windowing, as they achieve the smallest possible time-frequency spread (Rioul et al., 1991). In practice, other windows, such Hamming, Hann, are often used for engineering convenience or to control sidelobes, even though they do not achieve the minimal bound.

One may also interpret the STFT in terms of a filter-bank model, where each frequency bin corresponds to a bandpass filter obtained by modulating the same window $g(t)$. Specifically, a filter tuned to frequency $f_0$ has an impulse response proportional to $g(t)\,e^{-j\,2\pi f_0 t}$, so its bandwidth is determined by the spectral spread of $g(t)$. Since *all* these bandpass filters use the same window function, they share a common time spread $\Delta t$ and frequency spread $\Delta f$. This enforces *constant* resolution across the entire time-frequency plane, meaning that each filter provides the same temporal and spectral precision regardless of its center frequency. Consequently, if the window is narrowed to improve time resolution for one filter, it simultaneously broadens every filter's bandwidth, lowering frequency resolution everywhere.

Physically, $\Delta t$ and $\Delta f$ capture the STFT's resolution in time and frequency. If two events in the signal differ by less than $\Delta t$ in time or $\Delta f$ in frequency, the STFT representation typically merges them into a single indistinguishable feature (Rioul et al., 1991). By choosing a narrower window $g(t)$, one can sharpen time resolution but degrade frequency resolution. Conversely, broadening the window improves frequency discrimination at the expense of temporal clarity. Hence, $\Delta t$ and $\Delta f$ cannot be reduced simultaneously; narrowing one inevitably widens the other.

Different ways of quantifying spread may yield slightly different constants in the inequality, but the core principle remains the same: there is a lower bound on how sharply a signal can be localized in both time and frequency. This restriction explains why a single, fixed window in the STFT cannot simultaneously resolve short-lived transients and narrowly spaced frequency components across the entire time-frequency plane.

In many biomedical applications, signals might contain abrupt bursts at high frequencies and slowly changing trends at low frequencies. A uniform time-frequency resolution, dictated by a single window, becomes insufficient under these circumstances. While the trade-off between $\Delta t$ and $\Delta f$ persists, one may still seek an approach in which the window is effectively smaller at higher frequencies (enhancing time resolution) and larger at lower frequencies (enhancing frequency resolution). This leads to the idea of a *constant-Q filter bank*, where each filter's bandwidth is proportional to its center frequency (Rioul et al., 1991). Intuitively, rapid high-frequency bursts benefit from narrow time windows, whereas slow low-frequency fluctuations are better resolved with a broader window.

As we will see next, *wavelet transform* achieves this multi-resolution capability by varying the size of the analysis function – or *wavelet* – inversely with the frequency scale. The time-bandwidth product still adheres to Heisenberg's principle, but wavelet analysis allocates time-frequency resolution in a way that adapts more naturally to signals whose features unfold at different rates. This sets the stage for the Continuous Wavelet Transform, where scale-dependent wavelets can capture both fleeting transients and long-duration phenomena in a unified framework.

### IV.3.3 Continuous Wavelet Transform

Wavelet analysis extends the concepts of filtering and time-frequency decomposition introduced earlier, offering a more flexible framework for examining non-stationary signals. Recall from Section IV.2.2 that convolution with a suitably designed impulse response can emphasize or suppress particular signal components. Fourier-based methods similarly decompose a signal into sinusoidal basis functions, each extending infinitely in time (Fugal, 2009), so that abrupt changes in a non-stationary signal tend to spread broadly across the frequency axis in $X[k]$ (Rioul et al., 1991). By contrast, in the wavelet transform these unbounded sinusoids are replaced by small *wavelets* – localized functions that are (effectively) finite in duration and jointly supported in time and frequency.

Although wavelets can be viewed as convolution kernels, they differ from traditional filters in that their *scale* parameter changes to capture different frequency bands. Instead of using a fixed window length that slides over time as in the STFT, the wavelet transform contracts or expands the underlying *mother wavelet* according to the local frequency content (Cerutti et al., 2011; Rioul et al., 1991). Crucially, the time resolution is *not* uniform: shorter wavelets provide higher time resolution at higher frequencies, and longer wavelets yield higher frequency resolution at lower frequencies. This variable resolution aligns well with the needs of biomedical signal processing, where local bursts of high-frequency muscle activity, for example EMG spikes, and long-duration respiratory or cardiac oscillations often coexist.

From a conceptual standpoint, the *Continuous Wavelet Transform* (CWT) generalizes

this idea by treating both time and scale as continuous variables. If one instead treats the scale and translation parameters discretely, the wavelet expansion becomes a *wavelet series* (Rioul et al., 1991). In the continuous case, for a real- or complex-valued continuous signal $x(t)$, the CWT is defined in terms of *scaled* and *translated* versions of a chosen mother wavelet $\psi(t)$:

$$\psi_{a,b}(t) \;=\; \frac{1}{\sqrt{|a|}}\,\psi\Big(\frac{t-b}{a}\Big), \tag{IV.26}$$

where $a \neq 0$ is a *scale* parameter that compresses or stretches the wavelet in time, and $b$ is a *translation* or shift parameter that moves the wavelet along the time axis. Smaller values of $|a|$ produce narrower wavelets sensitive to higher-frequency content, whereas larger values of $|a|$ produce wider wavelets more attuned to lower-frequency behavior. The CWT of $x(t)$ with respect to $\psi$ is then defined as

$$\mathrm{CWT}_x(a,b) \;=\; \int_{-\infty}^{\infty} x(t)\,\overline{\psi_{a,b}(t)}\;dt, \tag{IV.27}$$

where the overbar denotes complex conjugation if $\psi$ is complex-valued. Conceptually, $\mathrm{CWT}_x(a,b)$ measures how strongly the localized wavelet $\psi_{a,b}$ matches the signal around time $t = b$. In other words, one obtains a "time-scale" representation in which fleeting high-frequency bursts can be captured by short wavelets, while more slowly varying dynamics are brought out by longer wavelets.

Because each scale corresponds roughly to an inverse frequency, the CWT can be viewed as a *bank of bandpass filters* whose center frequencies vary according to the parameter $a$. This interpretation ties naturally to the convolution-based filtering perspective from Section IV.2: for each scale $a$, the transform correlates the signal with a wavelet kernel that emphasizes a particular band of frequencies. The wavelet transform thus generalizes the STFT by letting the effective filter length, that is wavelet duration, shrink at higher frequencies and expand at lower frequencies, creating a true *multi-resolution* approach. Under appropriate conditions on the mother wavelet, one can also recover $x(t)$ from its CWT via an inverse transform:

$$x(t) \;=\; \frac{1}{C_\psi}\int_0^{\infty}\int_{-\infty}^{\infty} \mathrm{CWT}_x(a,b)\,\psi_{a,b}(t)\,\frac{db\,da}{a^2}, \tag{IV.28}$$

where $C_\psi$ is a constant that depends only on $\psi$ (Rioul et al., 1991). This ensures that wavelet-based analysis is a lossless transform, preserving signal energy in a way analogous to the energy-preservation properties encountered in the DFT and Parseval's relation.

The flexibility of the CWT makes it especially suitable for biomedical signals, where both rapid transients and slower oscillatory phenomena often arise in the same recording (Cerutti et al., 2011). For instance, brief bursts of high-frequency EMG activity would be

poorly resolved by a single wide analysis window, whereas a very short window might fail to capture low-frequency airflow fluctuations in OSA signals. By allowing the wavelet kernel to "adapt" across scales, one can retain fidelity to transient events and also reveal longer trends. Consequently, the CWT serves as a powerful extension of the classical filtering and STFT frameworks, enabling a deeper exploration of dynamic signal structure in both time and frequency.

As the above discussion shows, wavelets mitigate the STFT's uniform resolution issue by adapting the window to each frequency band in accordance with Heisenberg's uncertainty principle. However, the success of this multi-scale analysis hinges on the choice of the *mother wavelet* $\psi(t)$. Different mother wavelets exhibit different shapes in both the time and frequency domains, leading to varying degrees of localization and interpretability. In biomedical applications, selecting a mother wavelet that aligns with the physiological signal's morphology can significantly improve feature extraction.

### IV.3.4 Choice of Mother Wavelet

A *mother wavelet* lies at the heart of wavelet-based time-frequency analysis, dictating how signals are decomposed into localized oscillatory components. Different mother wavelets offer varying balances between time and frequency resolution, analytic properties, and sensitivity to transients or slow dynamics. The principal task is to select a family and parameterization that suit the specific nature of the signals under study, while respecting the uncertainty principle discussed in Section IV.3.2.

An *analytic wavelet* is a complex-valued function defined in continuous time that contains only positive-frequency components (Lilly et al., 2008). By having vanishing support on the negative-frequency axis, these wavelets prevent interference between positive and negative frequencies, leading to a clearer depiction of signal structure in the time-frequency plane. This separation of components simplifies amplitude-phase analysis and avoids ambiguities that might otherwise arise. Additionally, analytic wavelets tend to concentrate energy more effectively in the time-frequency plane, localizing around key events and thereby enhancing the detection of short-lived or transient phenomena.

A widely used analytic wavelet for biomedical applications is the *Morlet wavelet*. Conceptually, the Morlet wavelet is a modulated Gaussian function:

$$\psi_{\text{Morlet}}(z) \;=\; \pi^{-\frac{1}{4}}\, \mathrm{e}^{\mathrm{i}\,Q\,z}\, \mathrm{e}^{-\frac{z^2}{2}}, \quad Q > 0, \tag{IV.29}$$

where the parameter $Q$ influences the number of oscillations within its Gaussian envelope (Goupillaud et al., 1984; De Moortel et al., 2004). Because a Gaussian window satisfies the lower bound of the uncertainty principle, Morlet wavelets achieve near-optimal time-frequency localization, making them especially useful for signals with quasi-periodic

or oscillatory features. By adjusting $Q$, one can trade off frequency resolution against time localization: raising $Q$ narrows the frequency band while broadening the wavelet's time support, whereas lowering $Q$ sharpens time localization at the expense of coarser frequency selectivity.

Although Morlet wavelets excel at time-frequency localization, they require specific modifications to be strictly analytic (Lilly et al., 2008; De Moortel et al., 2004). In contrast, *generalized Morse wavelets* (GMWs) (Lilly et al., 2008) are exactly analytic by construction and can be viewed as a two-parameter family unifying multiple common wavelet forms. Let $\beta > 0$ and $\gamma > 0$ be the shaping parameters that govern GMWs' decay in both time and frequency domains. In the frequency domain, a GMW may be written as

$$\Psi_{\beta,\gamma}(\omega) \;=\; U(\omega)\, a_{\beta,\gamma}\, \omega^{\beta}\, \mathrm{e}^{-\omega^{\gamma}}, \tag{IV.30}$$

where $U(\omega)$ is the unit step (indicating positive-frequency support), $a_{\beta,\gamma}$ is a normalization constant, and $\omega$ is the angular frequency (Olhede et al., 2002; Lilly et al., 2012). By selecting specific $\beta$ and $\gamma$, one can obtain wavelets similar to well-known families such as the *Derivative-of-Gaussian* (DOG) wavelets when $\gamma = 2$, or approximate *Airy*-type wavelets when $\gamma = 3$. Because GMWs exclude negative frequencies, they simplify phase-related analysis and minimize interference artifacts in the time-frequency plane.

An important metric for characterizing GMWs is the *Heisenberg area*,

$$A_{\psi} \;=\; \sigma_t\, \sigma_{\omega}, \tag{IV.31}$$

where $\sigma_t$ and $\sigma_{\omega}$ are the standard deviations (spreads) of the wavelet in the time and frequency domains, respectively (Martinez-Ríos et al., 2022; Woyczynski et al., 2011). Heisenberg's uncertainty principle imposes

$$\sigma_t\, \sigma_{\omega} \;\;\geq\;\; \tfrac{1}{2}. \tag{IV.32}$$

In practice, raising $\beta$ for a given $\gamma$ increases the oscillatory content of the GMW, refining frequency resolution while reducing time localization. Conversely, lowering $\beta$ produces fewer oscillations and sharper peaks in the time domain (Lilly et al., 2012). As a result, the $(\beta, \gamma)$ parameter space provides a flexible handle on wavelet shape – one can "dial in" different time-frequency resolutions according to the physiological features of interest.

De Moortel et al. (2004) offers a valuable comparison of Morlet, Paul, and DOG wavelets, illustrating how each family addresses distinct analysis goals: Morlet wavelets generally yield robust frequency resolution but can smear closely spaced transient events, Paul wavelets excel at detecting abrupt changes in time but lack fine frequency discrimination, and DOG wavelets emphasize amplitude changes but may miss subtle phase information.

GMWs unify or approximate several of these wavelet forms, enabling to fine-tune wavelet shape for specialized tasks.

From a practical standpoint, a common approach is to evaluate multiple wavelet families and parameter choices, assessing their impact on classification or detection accuracy for the problem at hand (Martinez-Ríos et al., 2022). For signals featuring transient high-frequency bursts like alongside slower oscillations, an analytic wavelet like the Morlet or an appropriately parameterized GMW can illuminate both extremes effectively. If phase information is critical, analytic wavelets can significantly reduce distortion and interference. On the other hand, when detecting strong discontinuities or abrupt transients is important, DOG-like wavelets may prove advantageous.

In summary, the choice of mother wavelet – Morlet, GMW, or another family – influences how time-frequency features manifest in biomedical signals. Morlet wavelets, being near-Gaussian, have long served as a staple for capturing quasi-periodic phenomena in physiological measurements. Meanwhile, GMWs broaden these capabilities by offering exact analyticity and additional degrees of freedom through their parameters.

### IV.3.5 Scalogram

A *scalogram* provides a two-dimensional representation of a signal's energy in the time-scale plane, analogous to how the spectrogram visualizes energy in the time-frequency plane for the STFT. Recalling that the CWT of a signal $x(t)$ at scale $a$ and shift $b$ is given by the Equation IV.27 its *scalogram* is the squared magnitude

$$\text{Scalogram}(a, b) \;=\; \left|\text{CWT}_x(a, b)\right|^2. \tag{IV.33}$$

Plotting $\text{Scalogram}(a, b)$ with the time parameter $b$ on the horizontal axis and the scale parameter $a$ on the vertical axis (often inverted or converted to approximate frequency via $f \sim 1/a$) yields a 2D map of how the signal's energy is distributed across different scales and times. Color (or intensity) in the scalogram typically encodes the magnitude of $\left|\text{CWT}_x(a, b)\right|^2$, so bright or high-intensity regions in this plot signify strong correlations between $x(t)$ and a wavelet localized at that particular time and scale.

Much like the spectrogram, which is defined as the squared magnitude of the STFT, the scalogram displays a signal's energy in the time-frequency (or time-scale) plane (Rioul et al., 1991). However, unlike the STFT – which imposes a single, fixed window length for all frequencies – the wavelet transform operates at multiple resolutions. At smaller scales (high frequencies), the wavelets are briefer in duration, tightly localizing around a given time shift $b$. This property causes the influence of the signal's behavior at $t = t_0$ to be confined to a relatively narrow cone in the time-scale plane (Grossmann et al., 1990; Rioul et al., 1991). Conversely, at larger scales (low frequencies), the wavelets lengthen, capturing extended temporal contexts but reducing time precision. By comparison, any

localized event near $t = t_0$ in the STFT framework affects all frequencies within the fixed analysis window, making the region of influence equally wide at each frequency band.

Another distinction is the *logarithmic frequency progression* of wavelet scales. As the scale $a$ changes multiplicatively, the wavelet's central frequency also shifts roughly on a log scale. In contrast, the spectrogram typically samples frequencies on a linear grid, unless special filter banks are employed.

Fig. IV.4 illustrates this construction process step by step. In panel a, three wavelets of different scales are centered at a common reference time $t_0$, demonstrating the multi-resolution property of the CWT: smaller scales are brief and capture high pseudo-frequencies, whereas larger scales are elongated and emphasize low pseudo-frequencies. As each wavelet is slid along time, the resulting coefficients $\mathrm{CWT}_x(a, b)$ quantify how well the wavelet matches local features of the signal. Fixing a scale and plotting $|\mathrm{CWT}_x(a, b)|$ as a function of time yields the curves in panel b, where peaks occur at moments of strongest similarity between the wavelet and the signal. The width of each peak reflects the temporal support associated with that scale. Finally, by stacking these magnitudes across all scales and time shifts and squaring them to represent energy, one obtains the scalogram shown in panel c, in which bright regions indicate times and scales of high wavelet–signal correlation.

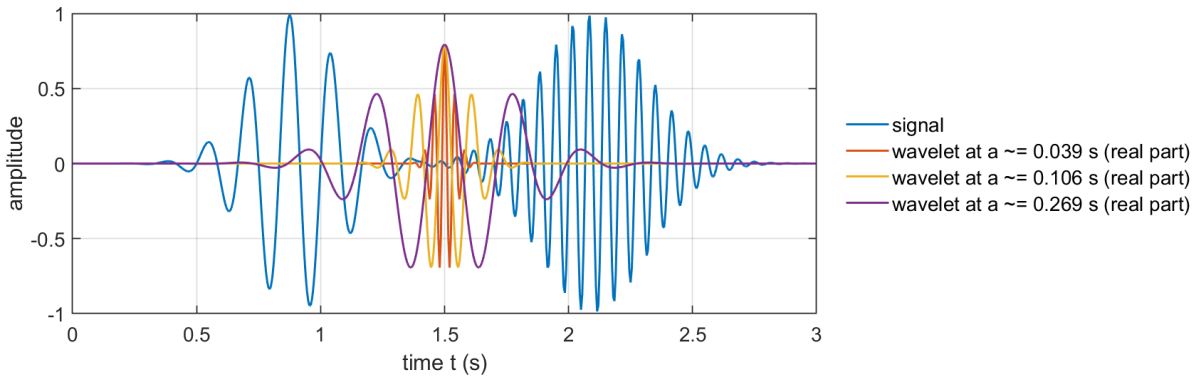
One noteworthy property of many wavelet transforms, under appropriate admissibility conditions on the mother wavelet $\psi$, is that they can preserve the energy of the signal up to a constant factor. In other words, integrating $\left|\mathrm{CWT}_x(a, b)\right|^2$ over time and scale can recover the total signal energy. A commonly cited formulation is (Grossmann et al., 1984; Rioul et al., 1991):

$$\int_0^\infty \int_{-\infty}^\infty |\mathrm{CWT}_x(a, b)|^2 \, \frac{\mathrm{d}b \, \mathrm{d}a}{a^2} = C_\psi \, \|x\|^2, \tag{IV.34}$$

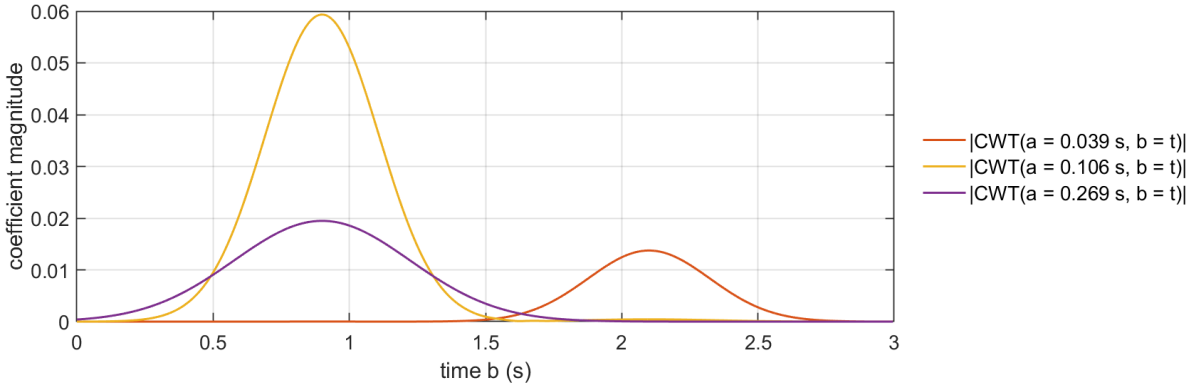
where $\|x\|^2$ is the energy of $x(t)$, and $C_\psi$ is a constant that depends only on the mother wavelet $\psi$. If $\psi$ is normalized so that $C_\psi = 1$, the wavelet transform is said to be *isometric*. This implies that the scalogram can serve as a bona fide energy distribution in the time-scale plane, analogous to how the squared STFT magnitudes act as an energy distribution in the time-frequency plane.

Despite its advantages for visualization, the scalogram also shares limitations with the spectrogram. A primary limitation is that the squared modulus of the transform, that is the energy distribution, *cannot, in general, be inverted* to recover the exact signal. Phase information is lost in the process of taking the modulus, and such phase details may be essential for accurate signal reconstruction. Additionally, because the scalogram (and spectrogram) are bilinear functions of $x(t)$, cross-terms may arise in the presence of

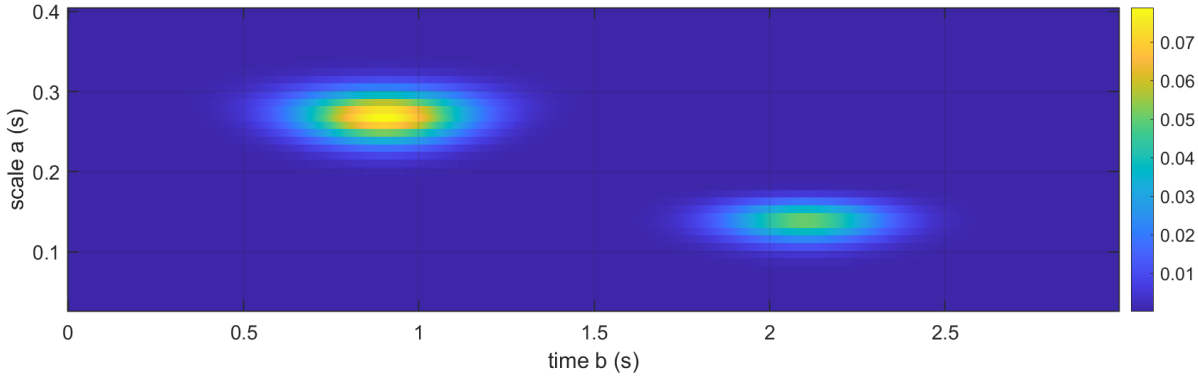
multiple overlapping patterns, manifesting as interference phenomena in the time-scale plane (Rioul et al., 1991; Kadambe et al., 1992).




**(a)** Sample signal with wavelets of different scales, centered at the same reference time $t_0$.




**(b)** CWT coefficient magnitudes $|\mathrm{CWT}(a, b)|$ at three fixed scales plotted as functions of time $b$.




**(c)** Scalogram $|\mathrm{CWT}(a, b)|^2$ over all times $b$ and scales $a$.

**Figure IV.4:** Construction of a scalogram of a sample signal: (a) choose a mother wavelet and inspect its scale-dependent time support at a common $t_0$; (b) slide each scale along time to compute $|\mathrm{CWT}(a, b)|$, whose peaks mark best matches; (c) arrange the squared magnitudes $|\mathrm{CWT}(a, b)|^2$ into a 2-D time–scale map whose bright regions summarize where energy concentrates across scales and time.

Although the scalogram highlights where the signal's energy is concentrated in time and

scale, certain localized transients may be more clearly revealed when amplitude *and* phase are considered (Grossmann et al., 1990; Rioul et al., 1991). In scenarios where subtle phase shifts are physiologically meaningful, particularly in biomedical signals with oscillatory and quasi-periodic patterns, analysis of the complex-valued CWT coefficients can be more informative than the scalogram alone.

Overall, the scalogram offers a rich visualization of how energy in a signal evolves across different time scales. Its multi-resolution nature provides an advantage over the spectrogram for analyzing signals that simultaneously contain rapid bursts and slower rhythmic patterns, a common occurrence in biomedical signals: for example, EMG spikes superimposed on lower-frequency respiration or cardiac oscillations. Despite the fact that the scalogram cannot directly be inverted to recover the original signal, it remains an invaluable tool for detecting features, identifying transient phenomena, and performing preliminary inspection of a signal's time-scale structure. In subsequent analyses, one may supplement scalogram insights with phase-based or full-complex CWT representations to obtain a more complete picture of non-stationary biomedical signals.

## IV.4 Image Processing

### IV.4.1 Images as a Two-Dimensional Signals and Their Basic Properties

Images can be viewed as two-dimensional signals that describe variations of a particular parameter over a spatial domain (Smith, 1997). In the most common case, this parameter is the intensity of visible light mapped onto a plane, resulting in what we usually call a *visual image*. Nonetheless, the concept of an image extends beyond the visible spectrum. Many other imaging modalities capture physical quantities, such as microwave or infrared radiation, temperature distributions, or even x-ray emissions. Once acquired, these measurements are typically transformed into grayscale or color images to facilitate human interpretation.

Similar to one-dimensional discrete-time signals, each element of a digital image is stored as a discrete sample. In imaging terminology, these samples are referred to as *pixels*. Each pixel typically encodes a numeric value representing the magnitude of the underlying measured quantity at a particular location in the original scene. For grayscale images, the intensity levels are commonly quantized into 256 discrete steps (ranging from 0 to 255), making them both manageable computationally and sufficiently detailed for the human visual system. In color imaging, three separate intensity values – red, green, and blue channels, are stored for each pixel, following the *RGB encoding* scheme. Allocating one byte to each of the three channels provides a potential range of $256 \times 256 \times 256 \approx 16.8$ million color combinations (Smith, 1997).

A key aspect of digital image enhancement involves modifying pixel values to improve the visibility of specific features. One common approach is the *grayscale transform*, which remaps the input brightness range to a new set of output levels. Although manual adjustments can selectively increase contrast in certain intensity intervals, this inevitably reduces contrast elsewhere. Automated methods such as *histogram equalization* reallocate intensities to emphasize frequently occurring brightness values, thereby enhancing the overall visibility of important regions (Smith, 1997). While these strategies are effective for global or regional image enhancement, they do not inherently capture or describe the subtle spatial arrangements of pixels that often define an image's finer details. For a deeper characterization of these local intensity patterns, known collectively as *texture*, we turn to specialized analysis methods. The following subsection introduces the fundamentals of *texture analysis* and discusses its significance in image processing.

### IV.4.2 Introduction to Texture Analysis

Texture analysis focuses on characterizing and quantifying the spatial organization of pixel intensities, often revealing subtle patterns that are imperceptible through simple brightness or contrast manipulations. Although no single universal definition of image texture exists, texture is generally viewed as a composite of repeated local patterns or "primitives" whose properties, such as brightness, color, size, and orientation, give rise to higher-level attributes like regularity, coarseness, and directionality (Rosenfeld, 1976; Levine, 1985; Materka et al., 1998).

Several main approaches exist for quantitative texture analysis. *Structural* methods describe texture in terms of well-defined primitives (microtexture) and the spatial arrangements (macrotexture) of these primitives (Haralick, 1979; Levine, 1985). Although such approaches can yield symbolic representations, they are difficult to apply when image structures are irregular and lack clearly defined primitives. By contrast, *statistical* methods do not require explicit models of repetitive elements; instead, they rely on statistical descriptors of pixel intensity distributions and inter-pixel relationships. In particular, second-order statistics derived from pairs of pixels have proven more effective than purely transform-based or structural methods in discriminating among textures (Weszka et al., 1976; Julesz, 1975). A key example within this class is the Gray-Level Co-occurrence Matrix, which provides a simple yet powerful way to capture how often pairs of intensity values co-occur at specified spatial offsets (Haralick, 1979).

*Model-based* methods employ fractal or stochastic processes to characterize texture by estimating the parameters of a generative or probabilistic model. Lastly, *transform-based* approaches, such as those using the Fourier or wavelet transform, represent texture in a frequency- or scale-related domain (Rosenfeld, 1976; Mallat, 1989; Laine et al., 1993; Lu et al., 1997). While Fourier transforms offer insight into periodicity, they are limited by their lack of spatial localization. Wavelets often address this limitation more effectively by

capturing local features at different scales. Each of these broad categories carries its own advantages and drawbacks, making the choice of method dependent on the nature of the texture and the application at hand.

In the next subsection, we focus on *statistical* texture analysis, and in particular the Gray-Level Co-occurrence Matrix. This method was chosen as a proof of concept in analyzing an OSA dataset because of its relatively straightforward implementation and proven utility in discriminating texture patterns.

### IV.4.3 Grey-Level Co-Occurrence Matrix

One of the classic statistical approaches to texture analysis is the *Gray-Level Co-occurrence Matrix* (GLCM) (Haralick, 1979). Its central idea is to measure how often pairs of pixels with particular intensity values appear in a given spatial relationship, determined by a specified distance and orientation. By focusing on these pairwise intensity transitions, the GLCM highlights *second-order statistics*, which capture relationships between pixel pairs rather than properties of individual pixels alone. This emphasis on joint intensity distributions makes the GLCM especially effective for characterizing textures that lack easily discernible structural elements or color differences (Haralick, 1979; Hung et al., 2019).

In practice, the GLCM is constructed by scanning through the image and recording the co-occurrence of pixel pairs at a chosen distance $d$ and orientation $\alpha$. For each pair of intensities $(i, j)$ found in that spatial configuration, one increments the matrix entry corresponding to $(i, j)$. After accumulating all such occurrences, the result is normalized to represent a joint probability. Formally, if $G$ is an image of size $M \times N$, and $(\Delta x, \Delta y)$ corresponds to the offset specified by $d$ and $\alpha$, the GLCM entry $p_{ij}$ is given by:

$$p_{ij} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}\big[G(m,n) = i \wedge G(m+\Delta x, n+\Delta y) = j\big]}{\text{Total number of valid pixel pairs}}, \tag{IV.35}$$

where $\mathbf{1}[\cdot]$ is an indicator function that equals 1 if its argument is true, and 0 otherwise. Because images often contain numerous intensity transitions, the GLCM can be computed for multiple orientations (e.g., $0°, 45°, 90°, 135°$) and distances, each yielding a distinct co-occurrence matrix. By registering how frequently specific gray-level transitions occur in localized spatial arrangements, this compact representation captures essential information about the underlying texture, without the need to explicitly identify or model individual texture elements.

Although the GLCM provides a structured representation of gray-level transitions, it is rarely used in its raw matrix form for classification or segmentation. Instead, descriptive numerical metrics, known as *GLCM features*, are computed to capture specific statistical properties of these transitions. These features include cluster tendency, contrast,

correlation, dissimilarity, entropy, homogeneity, maximum probability, and uniformity (also referred to as energy) (Haralick, 1979; Hung et al., 2019). Each feature emphasizes a distinct characteristic of the underlying texture.

One of the most common features is *contrast*, typically given by

$$\text{Contrast} \;=\; \sum_{i,j} (i-j)^2 \, p_{ij}, \tag{IV.36}$$

where $p_{ij}$ denotes the empirical probability of observing the intensity pair $(i, j)$ at the chosen spatial offset. This measure becomes larger in regions where neighboring pixels exhibit significant intensity differences. Another widely used feature is *energy (uniformity)*, defined as

$$\text{Energy} \;=\; \sum_{i,j} p_{ij}^2, \tag{IV.37}$$

which reaches higher values when the co-occurrence matrix is dominated by a small number of large probabilities, indicating a more uniform texture. In contrast, *entropy*,

$$\text{Entropy} \;=\; -\sum_{i,j} p_{ij} \, \log(p_{ij}), \tag{IV.38}$$

reflects the degree of randomness in the intensity transitions and thus attains larger values for more disordered textures.

*homogeneity* highlight transitions where the gray levels of neighboring pixels are similar.

$$\text{Homogeneity} \;=\; \sum_{i,j} \frac{p_{ij}}{1+|i-j|}, \tag{IV.39}$$

*Correlation* evaluates the linear dependence of intensities between neighboring pixels, and is defined as

$$\text{Correlation} \;=\; \frac{\sum_{i,j} (\, i-\mu_i \,)\,(j-\mu_j)\, p_{ij}}{\sigma_i \, \sigma_j}, \tag{IV.40}$$

where $\mu_i$ and $\mu_j$ are the mean intensities of the respective row and column, and $\sigma_i$ and $\sigma_j$ are their corresponding standard deviations. A higher correlation value indicates a stronger linear relationship between pairs of intensities across the image. Another complementary measure, *dissimilarity*, is typically written as

$$\text{Dissimilarity} \;=\; \sum_{i,j} |\, i-j \,|\, p_{ij}, \tag{IV.41}$$

where larger values signify greater differences between neighboring gray levels.

By condensing the co-occurrence matrix into these concise statistics, one can more reliably distinguish different textures without comparing entire GLCMs. Nonetheless, the effectiveness of GLCM-based methods is influenced by the choice of parameters such as the range of intensity values, distance offsets, and orientations used to compute co-occurrence. Despite these, the GLCM remains practical because of its conceptual simplicity and its ability to highlight second-order intensity relationships critical for texture-based discrimination.

## IV.5 Feature Extraction: Skeletal Muscle Disorders Datasets

In this section, we apply the concepts and methods introduced thus far to the skeletal muscle disorders dataset, with the ultimate aim of classifying EMG recordings into two pathological classes. As discussed in Subsection III.1.4, our procedure begins with feature extraction on the skeletal muscle channelopathy dataset. We also include the fibrillation potentials dataset to test whether the developed approach generalizes beyond channelopathies.

Before working with clinical recordings, we evaluated our methods on the synthetic dataset presented in Subsection III.1.2, which simulates both healthy and pathologically modified channels under controlled conditions. This dataset comprises three groups (healthy, sodium-channel defect, and chloride-channel defect), enabling us to verify the classification approach in a controlled setting. Because the synthetic signals are noise-free by design, there was no need to compute SNR or apply filtering.

In the subsequent sections, we describe the processing pipeline applied to the skeletal muscle channelopathy dataset, bearing in mind that most of these procedures were also used, when appropriate, on the fibrillation potentials dataset. The primary steps include plotting each EMG signal's PSD and spectrogram to evaluate signal quality, identifying dominant frequency bands, and computing SNR. We discuss when filtering is beneficial, measuring performance with metrics such as SNR, RMS error (RMSE), and cross-correlation.

We then extract both time-domain and frequency-domain features for benchmarking. Our emphasis is on creating feature sets amenable to advanced machine learning techniques, particularly convolutional neural networks that require two-dimensional image-like inputs. While spectrograms could fulfill this role, we opt for scalograms derived from the wavelet transform because of its adaptive frequency resolution, which is particularly suited to capturing transient myotonic discharges. Unless otherwise specified, all feature extraction in this section was performed in MATLAB (Inc., 2022).

### IV.5.1 Signal Quality and Processing

We first normalize our `.csv` files because different recordings may have varying amplitude ranges due to factors such as electrode placement, signal acquisition settings, or physiological differences among subjects. Normalization places all recordings on a consistent scale, which facilitates uniform processing across the dataset. Moreover, many signal processing algorithms perform better or converge more quickly on normalized data, as this prevents numerical issues arising from extremely large or small values. By adjusting each recording to a common reference point, normalization also ensures that comparisons of signal features between recordings are meaningful and unbiased.

Also referred to as *standardization*, z-score normalization transforms the data to have a mean of zero and a standard deviation of one. Given a data point $x$, the transformation is defined as:

$$z = \frac{x - \mu}{\sigma}, \tag{IV.42}$$

where $x$ is the original data point, $\mu$ is the mean of the dataset, and $\sigma$ is the standard deviation of the dataset.

We selected one of the highest-quality signals, `p59SD_R_RECT_FEM_1.csv` (belongs to sodium channel-defect class), based on auditory evaluation for its clarity and pronounced myotonic discharges. Fig. IV.5 shows the signal plotted in the time domain alongside its spectrogram.

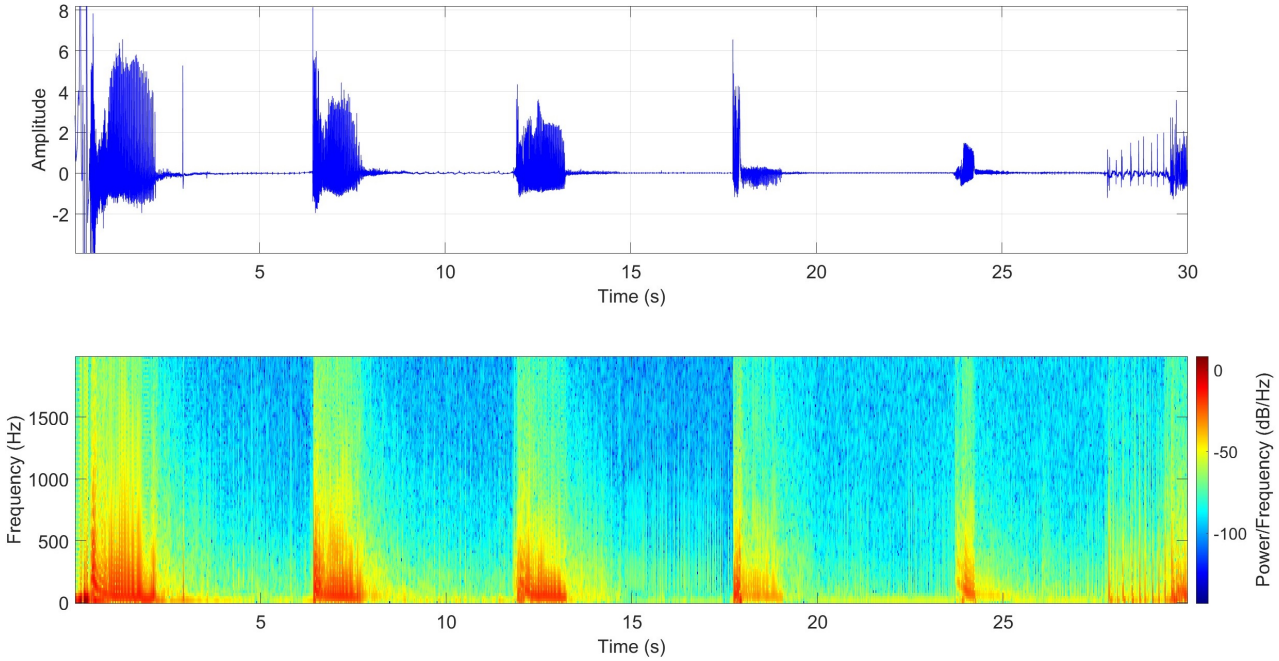



**Figure IV.5:** Time-domain plot of the iEMG signal (top) and its corresponding spectrogram (bottom).

This signal, which belongs to the sodium channel defect class, exhibits five distinct bursts of activity separated by intervals of relatively low amplitude (baseline): the first burst begins at approximately $t = 1\,\text{s}$. Subsequent bursts start at around $t = 6.6\,\text{s}$, $t = 12.2\,\text{s}$, $t = 18\,\text{s}$, and $t = 23.7\,\text{s}$. Each discharge typically lasts about 1–1.2 s, except the fifth, which is shorter, $\sim 0.6\,\text{s}$. These bursts correspond to myotonic discharges, while additional activity beginning around $t = 28\,\text{s}$ represents voluntary contraction trains.

At the start of the recording ($t < 1\,\text{s}$), there is a large spike caused by needle insertion. Smaller artifacts associated with needle movements also appear before certain discharges. As illustrated in Fig. IV.6, the high-amplitude signal at around $t = 6\,\text{s}$ precedes the second discharge, representing a brief needle adjustment used to trigger myotonic activity.

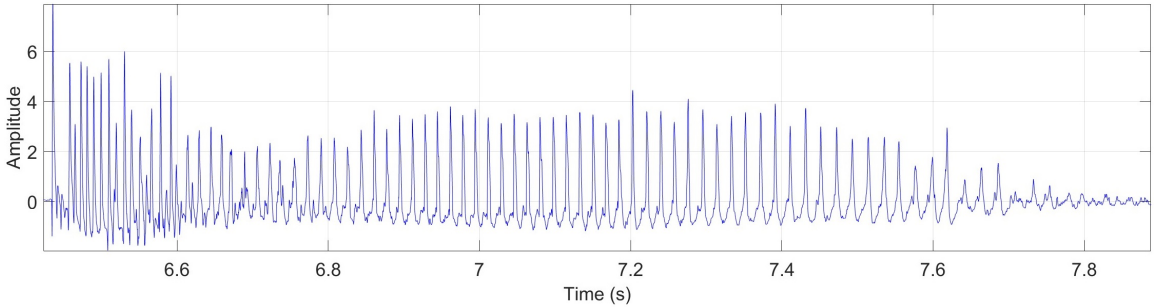



**Figure IV.6:** Close-up view of the second discharge in the iEMG signal. Note the large transient at $\sim 6\,\text{s}$ caused by needle movement prior to the discharge.

Between discharges, the signal remains relatively quiet, with low amplitudes observed during intervals from 5–6.6 s, 7.8–12.2 s, 13.2–18 s, and 19.1–23.7 s.

In the spectrogram of Fig. IV.5, *red* and *yellow* zones indicate higher power at certain frequencies, whereas *blue* regions reflect lower-power (mostly noise). The most intense band lies roughly below 200–300 Hz, implying that a considerable portion of the myotonic discharge energy resides in the low-to-mid frequency range. There is still noticeable power up to about 600–800 Hz (green/yellow), and faint energy may extend to 1 kHz or higher. By contrast, voluntary contractions appear to concentrate power at slightly lower frequencies than those dominated by myotonic discharges. No prominent single-frequency lines are visible, implying that simple notch filtering may not be necessary.

Because these discharges span a broad frequency range, there is substantial overlap between signal and noise across 0–300 Hz. Consequently, there is no trivial single band for extracting discharges without potentially cutting out relevant information. In addition, large amplitude artifacts like needle movements may artificially inflate the noise floor if included in baseline calculations. While a mild high-pass filter or a broad bandpass of, for example 20–1500 Hz, could potentially be helpful, narrow-band filtering risks eliminating crucial signal components.

To guide our subsequent processing decisions, such as choosing filter cutoff frequencies or excluding major artifacts, we next evaluated the SNR and computed PSD using Welch's method. We began by operationally defining the portions of the signal corresponding to discharge and those considered noise. In this context, the myotonic discharge activity is treated as the signal, while all other electrical activities, including voluntary contraction trains and needle-movement artifacts, are regarded as noise.

We identified the discharge intervals by both visual and auditory inspection, noting the specific time codes corresponding to each myotonic burst while labeling segments that contained needle-movement transients or strong voluntary activity as noise. After collecting all valid discharge segments into a single vector, we created a logical mask (`noiseMask`) initially set to `true`, implying that all samples were considered noise by default. We then excluded the identified discharge intervals by setting the corresponding indices in `noiseMask` to `false`, thereby isolating the noise-only samples.

Subsequently, we computed the RMS amplitude, representing average power, for both the discharge and noise portions. In this particular signal, the discharge RMS was 0.9895 whereas the noise RMS was 1.0021, leading to an SNR of -0.11 dB. Furthermore, the discharge exhibited its highest power at approximately 64.60 Hz, while the noise peaked near 21.53 Hz. Because large artifacts like needle movements can have amplitudes comparable to or exceeding those of the myotonic bursts, classifying these events as noise tended to inflate the noise power and thus resulted in negative SNR values.

We computed PSD with a frequency resolution of 23.53 Hz and a window duration of 23.2 ms, allowing us to compare the power distributions of discharge and noise over a broad frequency range. Fig. IV.7 shows the PSD of the sample recording, where the myotonic discharge generally exceeds the noise in certain mid-frequency bands but overlaps substantially in the lower and higher frequency regions. Overall, the PSD exhibits a typical $1/f$-type broadband profile, with power gradually decreasing as frequency increases. This extensive overlap between discharge and noise frequencies implies that no single band is exclusively associated with noise, making it risky to notch out any portion of the spectrum without potentially removing important discharge information.

As we plan to use scalograms for classification with machine-learning models, the resulting feature extraction should be sufficiently robust to accommodate the common bandwidth overlap, because the key waveforms distinguishing sodium versus chloride channel defects will remain. In contrast, needle movement artifacts, voluntary activity, and baseline noise are likely to appear similarly across all classes, leaving only the discharge patterns specific to each defect type as discriminative features. Consequently, filtering may not be strictly necessary if the classifier can learn to disregard shared noise frequencies. Nonetheless, we will also analyze a filtered data subset to evaluate the effect of filtering

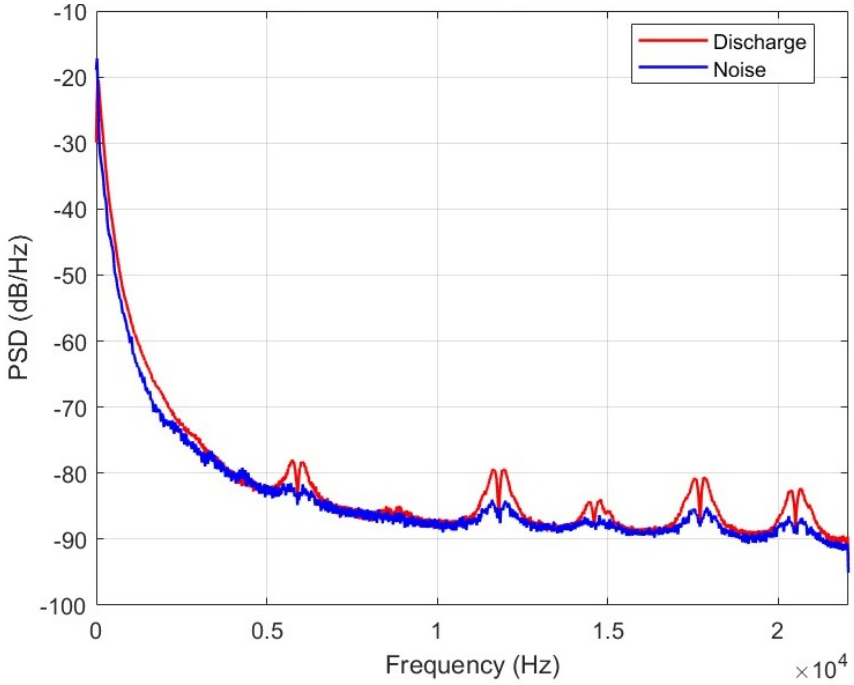



**Figure IV.7:** PSD comparison for discharge and noise, illustrating overlapping frequency bands.

on classification performance. Furthermore, manual SNR computation is challenging and, in our case, impractical due to the complex overlap of multiple signal components. By automating the process with a wavelet transform, we can generate scalograms that provide a comprehensive representation of the signal's complex structure and actual SNR, thereby eliminating the need for tedious manual signal decomposition.

We next illustrate potential data quality issues and the challenges of computing SNR and selecting filtering strategies by examining a second sample, `p54CR_L_BICEP_2.csv`, from the chloride channel-defect class. Although this recording is also considered high-quality—containing well-defined myotonic discharges and minimal overall contamination—it presents certain limitations relative to `p59SD_R_RECT_FEM_1.csv`. In particular, `p54CR_L_BICEP_2.csv` includes mild voluntary activity that frequently overlaps with the myotonic discharges, and some bursts occur in rapid succession or even simultaneously. These factors make it more difficult to isolate individual discharge events and thus complicate SNR estimation.

Fig. IV.8 shows the time-domain representation (top) and the corresponding spectrogram (bottom) of the `p54CR_L_BICEP_2.csv` signal. Several prominent bursts of high amplitude are interspersed with lower-level segments, while the largest peaks reflect needle movements, followed immediately by myotonic discharge trains. From the spectrogram, it is apparent that these bursts exhibit broadband energy primarily below 500–600 Hz, with additional power extending toward 1 kHz or higher. A background of mid-range activity

persists throughout, accompanied by intermittent spikes across multiple frequencies, likely representing overlapping sources such as intended muscle activity, involuntary discharges, and occasional artifacts. Overall, the spectrogram closely resembles that of `p59SD_R_RECT_FEM_1.csv`, suggesting that visual comparison of spectrograms alone might not suffice to distinguish sodium versus chloride channel defects.

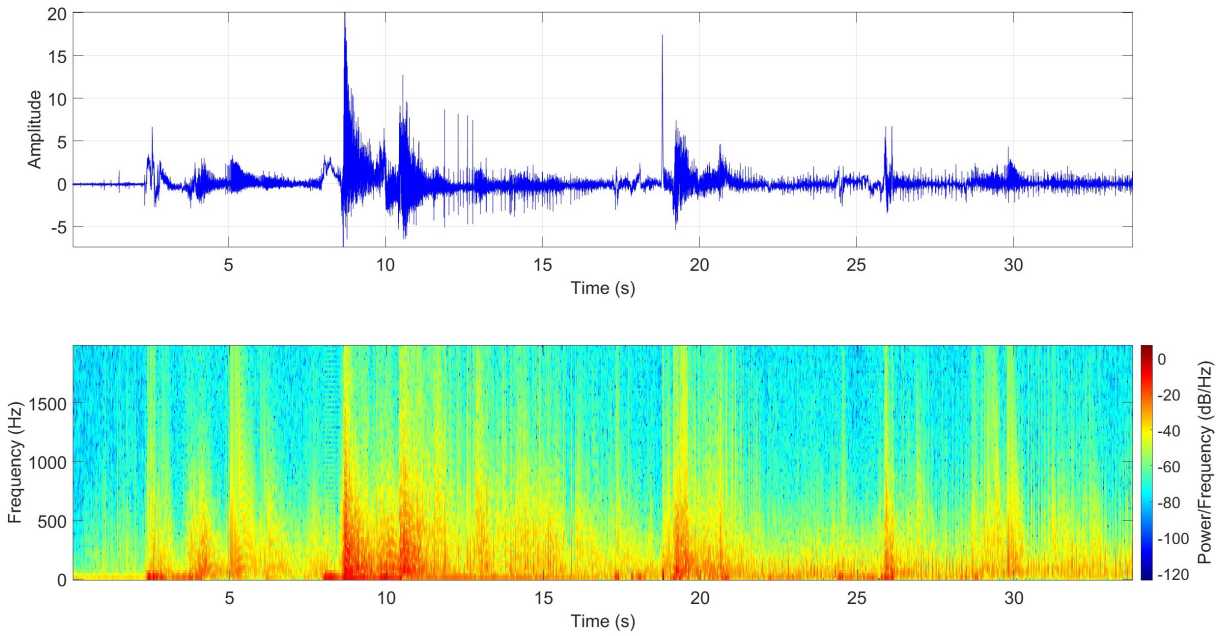



**Figure IV.8:** Time-domain plot (top) of the `p54CR_L_BICEP_2.csv` iEMG signal, along with its corresponding spectrogram (bottom).

A closer look at the signal segment beginning around $t = 19\,\text{s}$, shown in Fig. IV.9, following the large-amplitude needle movement artifact, provides an illustrative case. Although it is difficult to confirm from the plot alone precisely when myotonic discharges resume, auditory inspection of the recording indicates that previously initiated myotonic bursts are still audible at this point, and a newly triggered event begins around $t = 19\,\text{s}$, continuing until approximately $t = 22\,\text{s}$. Throughout this interval, voluntary activity remains in the background. Because multiple phenomena occur in parallel, accurately labeling particular segment as signal or noise becomes problematic and making straightforward SNR computations infeasible. This example shows the complexity of our samples where myotonic discharges, artifacts, and voluntary activity may overlap or occur in rapid sequence.

We began by computing individual PSD profiles for all signals within each channelopathy class and then averaged these PSDs to evaluate each group's overall spectral characteristics. We also plotted and evaluated spectrograms across the dataset to verify that the dominant frequency ranges remain consistent from recording to recording. These helped us select a suitable filtering range that would retain most of the myotonic discharge content while

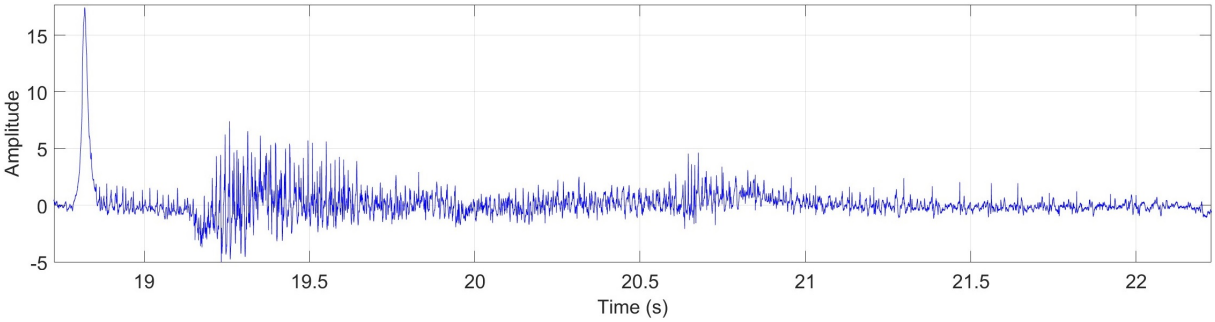
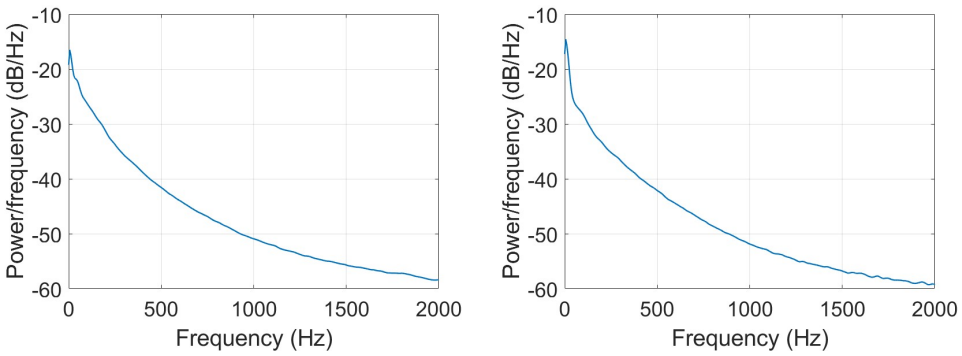



**Figure IV.9:** Close-up view of a train of events, including a newly triggered myotonic discharge in `p54CR_L_BICEP_2.csv`. The largest amplitude activity just before 19 s corresponds to needle movement, immediately followed by myotonic discharges.

removing undesirable signal content such as low-frequency drift and high-frequency noise.

Fig. IV.10 compares the averaged PSDs, computed in frequency range from 0 to 2000 Hz, for the sodium and chloride channel defect class datasets. In the sodium class, the peak frequency has a mean of 18.28 Hz (std = 22.6 Hz), whereas in the chloride class it averages 8.0 Hz (std = 15.21 Hz). Both curves exhibit a steep drop from near zero frequency up to a few hundred hertz. In the chloride class the spectrum begins near $-15\,\mathrm{dB\,Hz^{-1}}$ and drops rapidly, crossing $-30\,\mathrm{dB\,Hz^{-1}}$ by approximately 50 Hz. The sodium-channel defect class PSD follows the same overall $1/f$ contour but lies consistently $\sim 2$–$4\,\mathrm{dB\,Hz^{-1}}$ above the chloride spectrum from around 300 Hz to 1.5 kHz, indicating slightly stronger broadband activity. Thus, although the sodium spectrum is modestly elevated over much of the midband, both classes exhibit a smooth, monotonically decreasing profile without class-specific peaks or plateaus, suggesting that global PSD shape alone remains an insufficient discriminator.




**(a)** Sodium-channel defect class. **(b)** Chloride-channel defect class.

**Figure IV.10:** Averaged Welch PSDs for (a) the sodium channel defect class and (b) the chloride channel defect class.

Since both classes exhibit significant energy below approximately 300–500 Hz, yet also retain faint higher-frequency components up to about 1 kHz, we applied a mild second-order Butterworth band-pass filter from 20 Hz to 1500 Hz. This choice removes low-frequency drift while preserving the higher-frequency portion of the myotonic discharges observed in each dataset.

After filtering, signals belonging to the sodium channel defect class yielded an RMSE of mean = 0.565 (std = 0.257) and a maximum cross-correlation of mean = 0.760 (std = 0.207) when compared to their unfiltered counterparts. For the chloride channel defect class, the corresponding metrics were mean = 0.751 (std = 0.184) for RMSE and mean = 0.613 (std = 0.190) for the maximum cross-correlation. Although these values indicate moderate waveform changes, they also reflect a reasonable overall similarity, implying that the filter successfully attenuates noise and baseline drift without discarding key discharge features. Since neither averaged PSD displays strong peaks that would clearly distinguish the two classes, further time-localized spectral analyses will be necessary to isolate the characteristic bursts and repetition patterns for each defect type.

Since needle movements often produce the highest-amplitude peaks in the recordings, we selectively removed the two largest such events to observe how their absence might affect classification performance. In the SciPy package for Python (Virtanen et al., 2020), we employed the `find_peaks` function on both the raw signal and its negated version to detect positive and negative peaks that exceeded a specified amplitude threshold. We identified the two user-specified highest peaks from this combined set and replaced them via linear interpolation over a small window around each peak. This approach preserves the overall signal structure while eliminating the transient artifacts associated with needle adjustments. Finally, we saved the resulting, processed signals for subsequent analysis.

### IV.5.2 Time-domain Features

To build a comprehensive set of descriptors for the channelopathy dataset, we employed the EMG feature-extraction toolbox for MATLAB (Too et al., 2019), which provides both standard EMG measures and enhanced metrics designed to emphasize subtle waveform characteristics. Specifically, we extracted ZC, RMS, skewness, kurtosis, the mean absolute value (MAV), enhanced waveform length (EWL), log coefficient of variation (LCOV), slope sign changes (SSC), VAR, average energy (ME), temporal moment (TM), and maximum fractal length (MFL). Each of these features highlights a different aspect of the time-domain EMG signal. These features are computed on non-standardized signal in order to preserve absolute amplitude information.

A widely used amplitude-oriented feature, MAV is defined by taking the average of the

absolute values of each sample:

$$\text{MAV} \;=\; \frac{1}{L}\sum_{i=1}^{L}\big|\,x_i\big|, \tag{IV.43}$$

where $x_i$ are the EMG samples in a signal of length $L$. Due to its simplicity and effectiveness in various EMG classification tasks, MAV has become one of the most prevalent features (Too et al., 2019).

A wavelength (WL) measure (Hudgins et al., 1993), sums the absolute differences between consecutive samples, thus capturing waveform complexity by reflecting changes in slope:

$$\text{WL} \;=\; \sum_{i=1}^{L-1}\big|\,x_{i+1} - x_i\big|. \tag{IV.44}$$

The EWL further increases the contribution of mid-signal segments, where more physiologically meaningful changes often occur (Too et al., 2019). Formally,

$$\text{EWL} \;=\; \sum_{i=1}^{L-1} p_i\,\big|\,x_{i+1} - x_i\big|, \tag{IV.45}$$

where $p_i$ is a weighting function that boosts the influence of samples in the middle portion of the signal, allowing EWL to be more sensitive to subtle shape variations within an EMG burst.

LCOV measures relative fluctuations in the signal by comparing its standard deviation $\sigma$ to its mean $\mu$ in log space (Khushaba et al., 2017):

$$\text{LCOV} = \log\Big(\tfrac{\sigma}{\mu}\Big). \tag{IV.46}$$

This can reveal amplitude variability that might be obscured in simpler metrics such as the raw mean or standard deviation.

SSC counts how many times the slope changes sign, effectively capturing abrupt transitions in the signal's underlying MU activity. The general form can be written as

$$\text{SSC} \;=\; \sum_{i=2}^{L-1} f\big(x_{i-1}, x_i, x_{i+1}\big), \tag{IV.47}$$

where $f(\cdot)$ is an indicator function verifying that the slope between $x_{i-1}$ and $x_i$ differs in sign from the slope between $x_i$ and $x_{i+1}$.

Designed to capture waveform asymmetries, the TM integrates sample amplitudes weighted

by an integer order $k$. When $k = 3$, it highlights skewness in the amplitude distribution:

$$\mathrm{TM}_k \;=\; \sum_{i=1}^{L} i^k \, x_i, \tag{IV.48}$$

which can detect subtle imbalances in myotonic bursts (Hudgins et al., 1993).

Finally, MFL estimates the fractal dimension of the EMG waveform, quantifying its irregularity or complexity (Phinyomark et al., 2012):

$$\mathrm{MFL} \;=\; \max\Big\{\, \mathcal{F}(\{x_i\}) \,\Big\}, \tag{IV.49}$$

where $\mathcal{F}$ is a fractal estimation function evaluated over the signal $\{x_i\}$.

These features were chosen to balance amplitude-sensitive descriptors like EWL, LCOV with measures of waveform complexity (SSC, MFL) and distributional asymmetry (TM). By combining amplitude-, complexity-, and morphology-based metrics, we aim to highlight subtle distinctions in pathological muscle activity that may arise from sodium versus chloride channel defects, ultimately improving the robustness of our classification system.

### IV.5.3 Features for Image-based Learning

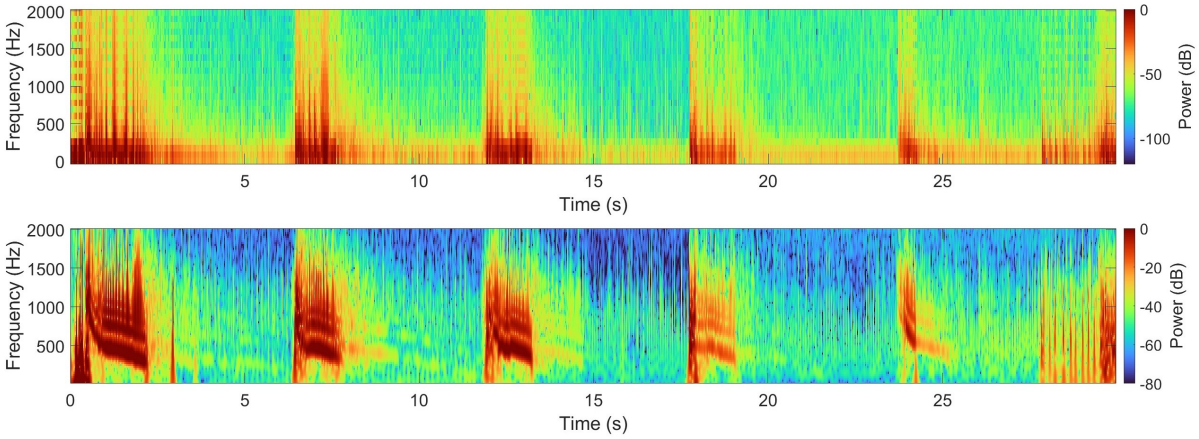



**Figure IV.11:** Spectrogram plot of the EMG signal (top), scalogram obtained using the Morse wavelet (bottom) of the signal `p59SD_R_RECT_FEM_1.csv`. The scalogram displays finer resolution for transient events, capturing short, high-frequency oscillations and gradually waning discharges more distinctly than the spectrogram.

To exploit advanced machine learning methods that require rich feature representations, we transformed each iEMG signal into a 2D time-frequency representation, a scalogram, using the CWT. As illustrated in Fig. IV.11, comparing the scalogram with a conventional

spectrogram immediately reveals sharper transitions and clearer delineation of dominant frequency bands, especially for short bursts or rapidly decaying oscillations typical of myotonic discharges. Unlike the spectrogram's fixed-window approach, which can blur brief high-frequency phenomena, the CWT adapts to local scale changes, yielding finer resolution for transient events within the targeted frequency range (here up to 2 kHz). As a result, oscillations whose frequency and amplitude diminish over a short interval appear distinctly in the scalogram but may be partially masked in the spectrogram. This capacity to highlight subtle, higher-frequency details or short-lived bursts makes the wavelet-based representation more effective at revealing channelopathy-specific patterns that might otherwise remain hidden.

We explored both the Morlet and Morse mother wavelets to generate our scalograms. The Morlet wavelet efficiently captures oscillatory events and can be tuned by adjusting the number of voices per octave and its scale-to-frequency mappings. The Morse wavelet, by contrast, provides additional flexibility via two parameters: $\gamma$ (the symmetry factor) and $P$ (the time–bandwidth product). Setting $\gamma = 3$ ensures zero skewness, minimizing the wavelet's Heisenberg area and yielding optimal localization in both time and frequency. Increasing $P$ prolongs the wavelet's effective duration, thereby improving resolution for slower, low-frequency discharges. Since our data concentrate power mainly below a few hundred hertz, we also specified more voices per octave to achieve finer granularity in this critical range. Depending on our experimental goals, we either covered a broad bandwidth for benchmarking or targeted 20–1000 Hz to exclude very low-frequency drift and high-frequency noise.

To expand training diversity, we applied small temporal shifts to the signals, producing augmented samples that capture minor phase and onset variations in myotonic discharges. Once we computed the absolute values of the CWT coefficients at each time–frequency bin, we normalized them and applied a color map (jet) to form a scalogram. Finally, we resized each scalogram to either $224 \times 224$, $299 \times 299$, $331 \times 331$, $384 \times 384$ pixels, matching the input dimensions expected by different convolutional neural network (CNN) architectures. This conversion from one-dimensional signals to colorized time–frequency images preserves the essential amplitude and frequency patterns of the iEMG data, while enabling standard image-classification pipelines to detect subtle channelopathy-specific features.

As a result of the wavelet-based processing, we obtained principal dataset variants, determined by the mother wavelet type (Morlet or Morse), frequency ranges, wavelet parameters and time-frequency resolution trade-off (voices per octave). This strategy allows us to assess whether different wavelet configurations influence the network's ability to distinguish sodium channel defects from chloride channel defects.

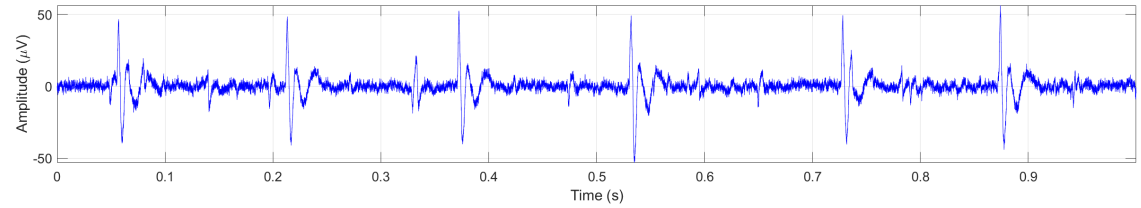
### IV.5.4 Synthetic Skeletal Muscle Channelopathy Dataset

This dataset consists of simulated iEMG signals designed to reflect ion-channel defects under controlled, well-defined conditions. Because the signals are generated synthetically, no additional filtering or artifact removal was necessary. Instead, we used these recordings primarily as a proof of concept to test the feasibility of our wavelet-based feature extraction pipeline for channelopathy classification. Following the same CWT procedure described previously, we created scalograms using the Morse wavelet with 12 voices per octave, $\gamma = 3$, and $P = 60$. These parameters produced time–frequency representations for three categories of simulated subjects: healthy controls, sodium-channel defect models, and chloride-channel defect models. By including healthy simulated subjects, we ensured that our framework could detect potentially subtle or intermediate changes indicative of channelopathies, thus validating the general approach under precisely known (simulated) conditions.

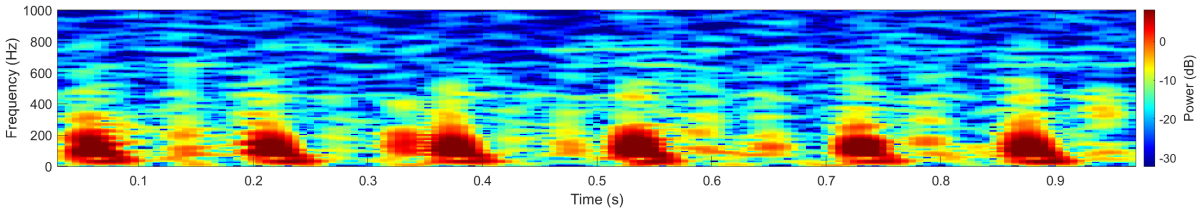
### IV.5.5 Skeletal Muscle Fibrillation Potentials Dataset

While working with the fibrillation potentials dataset, our aim was not so much to benchmark a range of transformations as to verify the universal applicability of the final classification algorithm. Consequently, we opted for the wavelet configuration that had proven most robust in earlier experiments: the Morse wavelet with 12 voices per octave, $\gamma = 3$, and $P = 60$. Scalograms were generated in the same fashion described for the channelopathy datasets, but without further exploration of alternative wavelets or parameter grids. By consistently applying the best-performing parameters identified in our prior analyses, we tested whether the learned classifiers could also detect fibrillatory characteristics beyond typical myotonic or normal muscle behavior. This approach illustrates how the chosen scalogram creation process, once tuned, can generalize across diverse EMG pathologies.

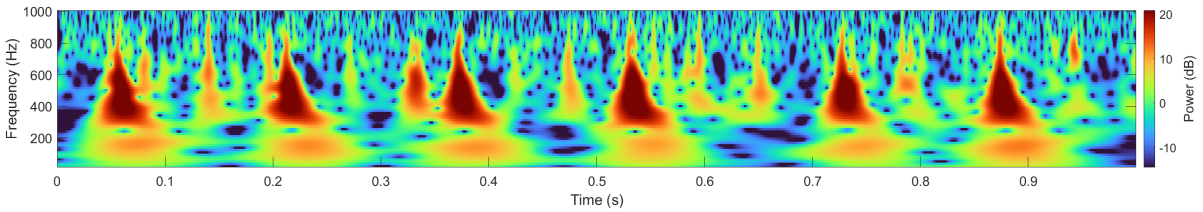
The characteristic differences between fibrillation and voluntary action potential recordings are illustrated in Fig. IV.12 and IV.13, respectively. In the time domain (Figures IV.12a and IV.13a), fibrillation potentials appear as dense trains of short, sharp spikes with quasi-regular firing, whereas voluntary motor-unit action potentials are fewer, longer, and poly-phasic. In the corresponding spectrograms (Figures IV.12b and IV.13b), fibrillation shows repeated broadband bursts extending to higher frequencies, while voluntary activity concentrates energy below approximately 300–400 Hz. The CWT scalograms (Figures IV.12c and IV.13c) further emphasize these contrasts: fibrillation exhibits numerous tall, narrow ridges spanning the upper frequency range, while voluntary activity yields fewer, broader, low-frequency cones. Collectively, these representations highlight the impulsive, high-frequency nature of fibrillation potentials versus the sustained, low-frequency pattern of voluntary activation.




**(a)** Time-domain representation of the signal.




**(b)** Spectrogram.



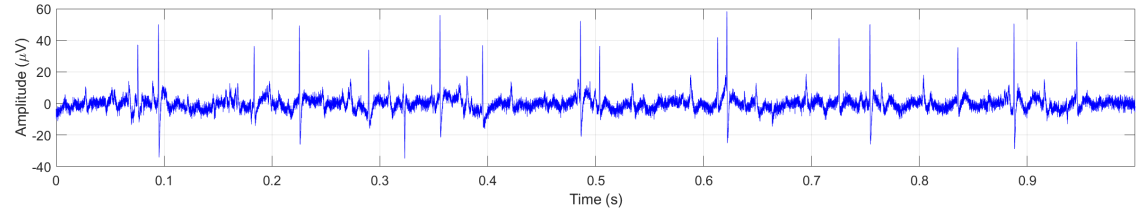

**(c)** Scalogram.

**Figure IV.12:** Time-amplitude and time-frequency domain representations of a 1-second segment of the voluntary action potentials trains iEMG recorded from patient 19.

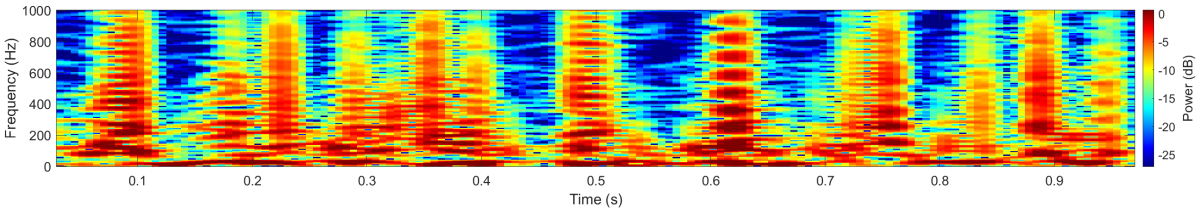
## IV.6 Feature Extraction: Obstructive Sleep Apnea Dataset

In this section, we focus on extracting meaningful features from patient airflow signals, as described in Subsection III.2.1 for classifying responders versus non-responders to MAS treatment. Most recordings are acquired at a sampling rate of 10 Hz, which, according to the Nyquist criterion, limits spectral analysis to frequencies up to 5 Hz. In contrast, two data subsets – PhysMAS and CRC, are recorded at higher rates, up to 250 Hz, theoretically allowing examination of frequencies up to 125 Hz. Although lower-frequency components generally provide the most informative physiological content for normal breathing, in pathological cases the assessment of higher-frequency elements may be critical to solving this classification problem.

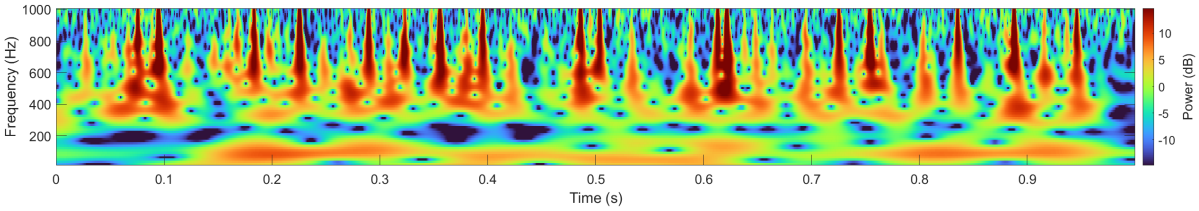
The extended duration of the recordings, typically spanning 7 to 9 hours, poses significant computational challenges. In addition, the non-stationary nature of the signals, arising from variations in respiratory events throughout the night complicates the analysis. Ensuring consistent feature extraction across recordings with different lengths and sampling rates is therefore essential. To address these issues and enable a meaningful




**(a)** Time-domain representation of the signal.




**(b)** Spectrogram.




**(c)** Scalogram.

**Figure IV.13:** Time-amplitude and time-frequency domain representations of a 1-second segment of the fibrillation potential iEMG recorded from patient 68.

comparison between MAS responders and non-responders, a systematic preprocessing framework is required.

The feature extraction process, developed using MATLAB (Inc., 2022), unless otherwise specified, begins with data loading and inspection to verify signal quality and identify any sampling frequency discrepancies. Next, the signals undergo preprocessing that includes resampling and standardizing the recordings. The long-duration signals are then segmented into manageable epochs, allowing the assumption of quasi-stationarity within each segment for reliable spectral analysis. Time-frequency analyses are performed using methods such as Welch's power spectral density estimation, spectrograms, and scalograms, which together provide a comprehensive visualization of the frequency content. Finally, a range of time-domain, spectral, and image texture features is extracted from signals, and these features are aggregated for each class to facilitate subsequent classification.

### IV.6.1 Signal Quality and Features for Non-image based learning

In this section, we describe the preprocessing and feature extraction procedures applied to the OSA datasets. The datasets under study – OSAMAS, CRC, and PhysMAS, exhibit substantial variations in both sampling frequency and recording duration, as described in Section III.2. Consequently, the range of reliably observable frequencies

differs significantly among the datasets. For instance, most OSAMAS signals are acquired at 10 Hz, which restricts the maximum observable frequency to 5 Hz according to the Nyquist criterion, whereas PhysMAS signals, sampled at 250 Hz, theoretically permit analysis of frequencies up to 125 Hz. Although the most informative spectral content of airflow signals generally resides in the lower-frequency range, in pathological cases such as OSA the higher frequency components may be critical for evaluating the effectiveness of MAS therapy, as OSA events might have significant high-frequency component. While implementing the preprocessing steps, we tried to ensure that the feature extraction steps will preserve the essential characteristics needed to distinguish MAS responders from non-responders.

The processing begins by loading the `.csv` files containing the polysomnography-derived airflow signals from each patient. Each dataset is processed separately. The sampling frequency is estimated automatically from the timestamp information contained in the `.csv` files. When the original sampling frequency exceeds the target rate (the most prevalent sampling rate in the dataset), the signals are downsampled to a common rate so that direct comparisons can be made between recordings within the same dataset. This downsampling guarantees that subsequent spectral computations are performed on signals with uniform temporal resolution.

In each dataset, the recording with the shortest duration is first identified, and all recordings are then standardized to a common length to facilitate further classification. For instance, while most recordings in the OSAMAS and CRC datasets span between 7 and 9 hours, the shortest recordings are approximately 4.99 hours and 6.31 hours long, respectively. To ensure consistency and reduce computational complexity, we standardize these datasets to a fixed length of 4.5 hours. Similarly, for the PhysMAS dataset, where the shortest recording is 1.82 hours, all signals are standardized to 1.8 hours, which is equivalent to 108 minutes. For further analysis, the standardized signals are segmented into 1-minute epochs. This segmentation not only reduces the computational burden but also allows us to assume quasi-stationarity within each epoch, which is critical for reliable spectral analysis. PSDs of both the full-length signals and the shortened segments are computed using Welch's method with a window length of one minute and a 50 % overlap, ensuring robust estimation of the frequency content.

In parallel with the spectral analysis, time-domain features, such as mean, variance, std, RMS, skewness, kurtosis, ZC, LCOV, SSC, and TM are computed for each patient. In order to extract more refined spectral descriptors from the full-length airflow signals, we determine the five dominant frequencies within each of three frequency bands (low, mid, and high). For the OSAMAS dataset, the default frequency ranges are defined as 0.1–2.0 Hz for the low-frequency band, 2.0–3.0 Hz for the mid-frequency band, and 3.0–5.0 Hz for the high-frequency band; for the CRC and PhysMAS datasets, these ranges

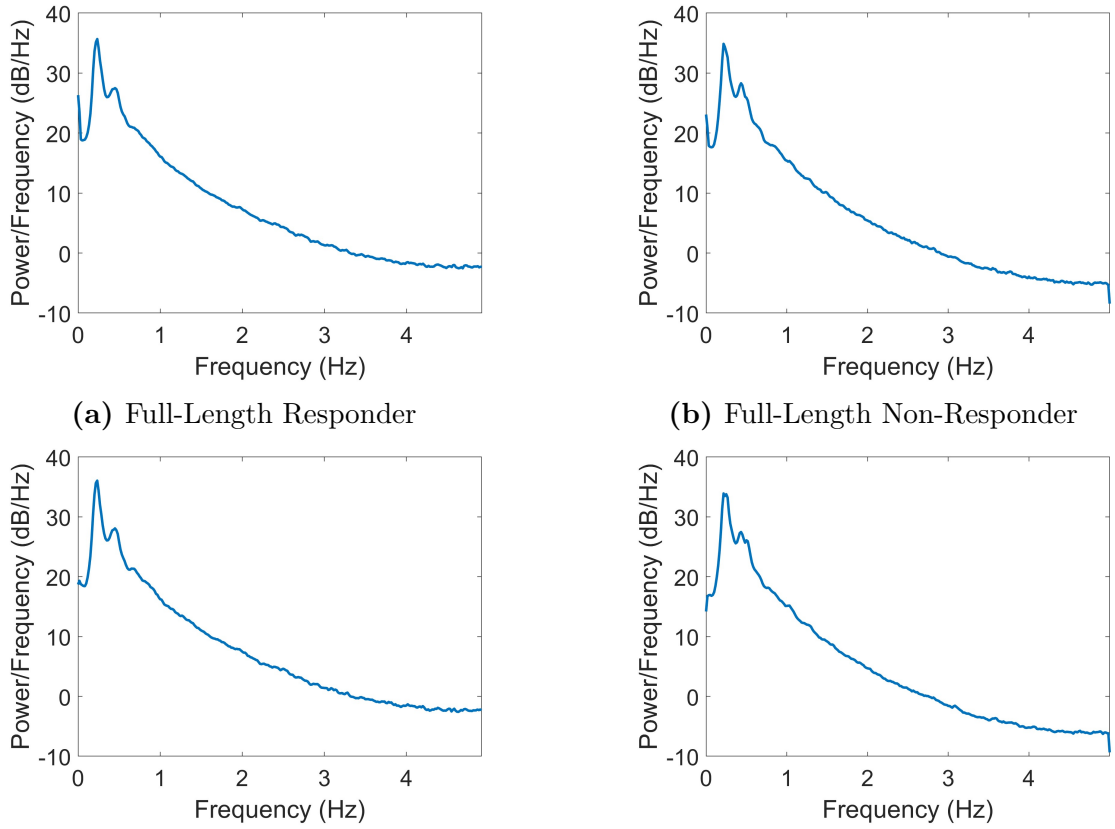
were adjusted to reflect their higher sampling rates and expanded spectral content as 0.1–2.0 Hz for the low-frequency band, 2.0–10.0 Hz for the mid-frequency band, and 10.0–50.0 Hz (125.0 Hz for PhysMAS) for the high-frequency band. All extracted features are then aggregated separately for MAS responders and non-responders, identified by the patient folder labels facilitating a clear comparative analysis.

The outputs of this processing pipeline include individual PSD plots for both the shortened and the full-length signals, averaged PSD plots, obtained by averaging the individual PSDs computed separately for responders and non-responders, and `.csv` tables summarizing basic statistics and extended time-domain features along with dominant frequency values for each patient. These comprehensive results enable us to assess signal quality and feature consistency across the three datasets.

Fig. IV.14 presents the averaged PSDs for the OSAMAS dataset, computed from both full-length and shortened signals for responders and non-responders. The overall spectral profiles of the signals remain consistent regardless of whether full-length or truncated segments are analyzed. In both cases, the primary spectral peaks, typically observed around 0.2–0.3 Hz and 0.6–0.7 Hz, are preserved, indicating that key physiological fluctuations are maintained after signal truncation. A modest elevation in low-frequency power is apparent for responders, though the differences between the classes are subtle and warrant further statistical validation.

In contrast, the CRC dataset (see Fig. IV.15) exhibits a less steep decline in power and a relatively flatter profile across the low and mid frequencies, suggesting that low-frequency fluctuations are less pronounced. Across all conditions – full-length versus shortened recordings and responders versus non-responders – the PSDs display a predominantly 1/f-like decay, with power starting at a relatively high level near the lowest frequencies and steadily decreasing to approximately −50 dB by 50 Hz. The overall shape remains consistent between full-length and truncated data, indicating that truncation does not substantially alter the broad distribution of spectral power. While subtle variations in power across frequency bands between responders and non-responders are apparent, particularly in the mid-frequency range (approximately 10–20 Hz), these differences are relatively minor and need more detailed formal analysis.

In the PhysMAS dataset (see Fig. IV.16), both the full-length and the shortened airflow signals exhibit a steep initial drop in power from near 0 Hz to approximately 10 Hz, followed by a more gradual decline across the mid-frequency range and a smooth tapering that extends just beyond 120 Hz. Shortening the signals does not fundamentally alter the overall spectral shape; the key features – an early steep slope, a mid-frequency plateau, and a gradual high-frequency taper, are consistently preserved in the averaged PSDs. Although the spectral profiles for responders and non-responders in the PhysMAS dataset




**(a)** Full-Length Responder

**(b)** Full-Length Non-Responder

**(c)** Shortened Responder

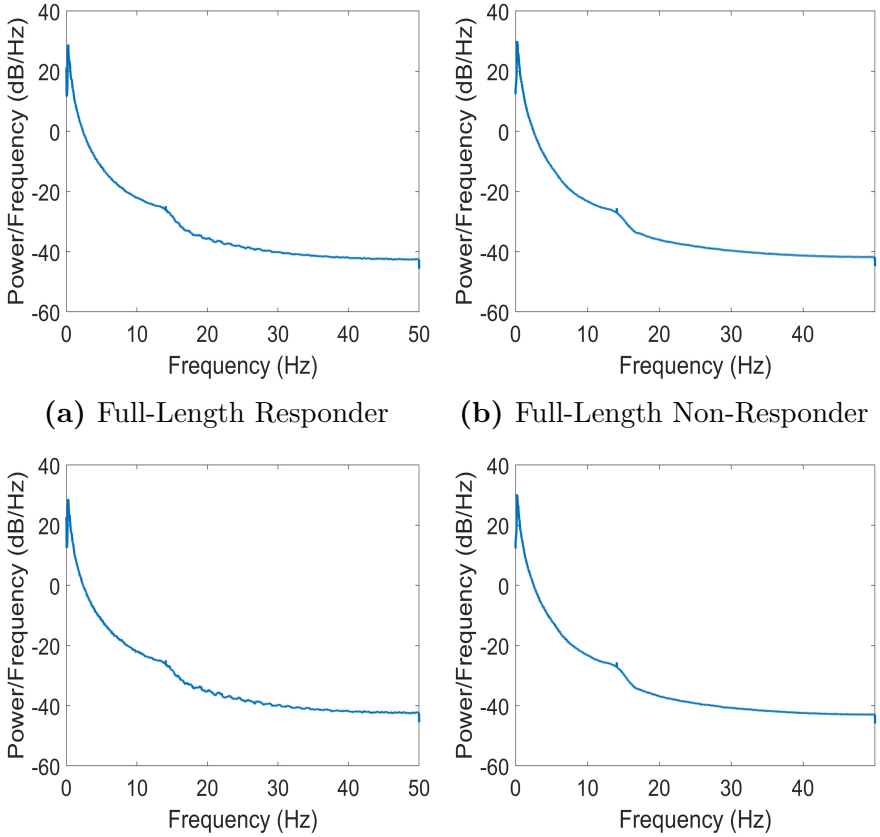
**(d)** Shortened Non-Responder

**Figure IV.14:** Comparison of PSDs for patient airflow in OSAMAS dataset: full-length signals (top) and shortened signals (bottom) for responder and non-responder classes.

share a similar overall shape, subtle differences are evident. In particular, there are slight shifts in mid-frequency power and a marginally deeper dip at higher frequencies for non-responders. These group-level distinctions may hint at underlying physiological differences; however, more rigorous analysis is needed to establish their significance.

When comparing all three datasets, a general decrease in power with increasing frequency is observed, but the slope and behavior in the mid-frequency range vary in subtle ways. For the OSAMAS dataset, the low sampling rate restricts the analysis to frequencies up to 5 Hz, precluding evaluation of higher-frequency content. In contrast, the CRC and PhysMAS datasets retain substantial energy beyond 5 Hz. Moreover, differences in the PSD curve shapes for responders and non-responders become evident at frequencies above 10 Hz in the CRC and PhysMAS datasets, while the OSAMAS PSDs, spanning only the 0–5 Hz range, appear nearly identical for both classes. These observations suggest that mid- and high-frequency content may serve as crucial distinguishing characteristics for classification.

To further study the effects of a lower sampling rate, we compared two signal segments




**(a)** Full-Length Responder **(b)** Full-Length Non-Responder

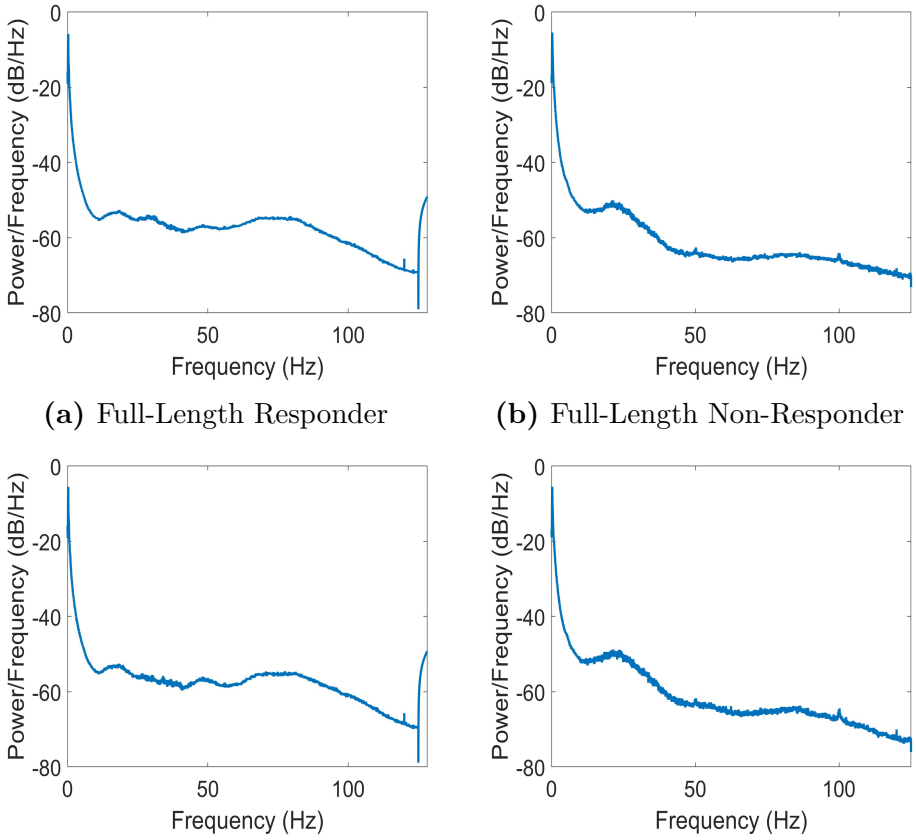
**(c)** Shortened Responder **(d)** Shortened Non-Responder

**Figure IV.15:** Comparison of PSDs for patient airflow in CRC dataset: full-length signals (top) and shortened signals (bottom) for responder and non-responder classes.

containing respiratory events from the OSAMAS dataset. One segment was taken from OSAMAS_012, which has the highest sampling rate in the dataset (256 Hz), while the other was taken from OSAMAS_029, recorded at a typical sampling frequency of 10 Hz. In each case, we identified a single hypopnea event and created an extended signal by adding 30 seconds of data before and after the event. We then plotted these extended segments and derived their PSDs, spectrograms, and scalograms, as shown in Fig. IV.17.

In the time-domain plot, Fig. IV.17a, a pronounced reduction in airflow amplitude is highlighted in pink, marking a hypopnea episode; shortly afterward, a light-blue section denotes a brief arousal, where the amplitude surges before settling back into a more stable pattern. The PSD shown in Fig. IV.17c confirms that most of the signal's energy is concentrated in the lower-frequency range, consistent with typical breathing rates with a steep drop-off in power at higher frequencies. However, a slight increase in power is observed starting at 60 Hz. This trend is also evident in the spectrogram (Fig. IV.17e ), where increased spectral power in the 60–100 Hz range is apparent during the hypopnea event, followed by a reduction toward its end and a subsequent surge coinciding with the arousal and restoration of normal airflow amplitude. The scalogram in Fig. IV.17g mirrors

the spectrogram's pattern with higher resolution: the red and orange bands at lower frequencies become attenuated during hypopnea and then regain intensity around the time of arousal. Notably, the faint high-frequency region (above 100 Hz) observed during hypopnea in the spectrogram becomes more pronounced in the scalogram, emphasizing a temporary shift of spectral power into very high-frequency regions during the respiratory event, which then disappears as normal breathing resumes.




**(a)** Full-Length Responder **(b)** Full-Length Non-Responder

**(c)** Shortened Responder **(d)** Shortened Non-Responder

**Figure IV.16:** Comparison of PSDs for patient airflow in PhysMAS dataset: full-length signals (top) and shortened signals (bottom) for responder and non-responder classes.

Compared to signal OSAMAS_012, which was sampled more densely and exhibited high-frequency energy extending above 60 Hz, OSAMAS_029 recording plotted in Fig. IV.17b appears noticeably "smoothed out." Its PSD shown in Fig. IV.17d, is concentrated below 5 Hz and lacks the broad high-frequency content present in OSAMAS_012; this absence is similarly reflected in the spectrogram (Fig. IV.17f). As it can be seen in Fig. IV.17h, the scalogram no longer displays the fine-grained transitions at higher frequencies, the subtle shift of dominant frequency bands to higher frequencies and the fading of spectral power in the low- and mid-frequency ranges during hypopnea. Although OSAMAS_029 captures the essential breathing rhythm, it fails to resolve the nuanced flow dynamics evident in the higher-frequency range of OSAMAS_012. This discrepancy is due to the

lower sampling rate of OSAMAS_029, which constrains its overall bandwidth and resolution.

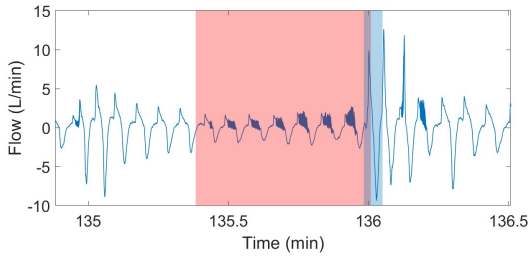
Based on these findings, the next steps will involve a statistical evaluation of the extracted features to identify those that best discriminate between MAS treatment responders and non-responders. In the next chapter, we will then assess whether classification models can be trained on the combined datasets or if dataset-specific models yield superior performance. Comparing the results of separate and compound evaluations of the datasets might help us to find out whether the hight frequency component of the signals is a decisive discriminatory characteristic. If this is the case, then classification metrics of OSAMAS dataset alone should be worse than that of CRC and PhysMAS. This preprocessing framework, therefore, not only addresses the challenges posed by variable sampling rates and long recording durations but tries to systematically explore the signal characteristics most relevant to OSA classification.

### IV.6.2 Features for Image-based Learning

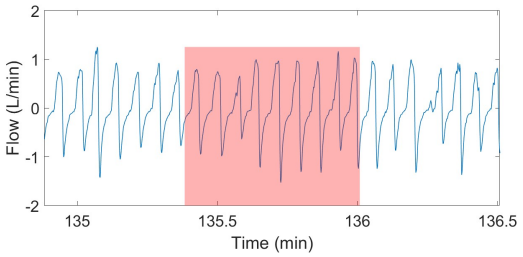
We now extract signal features for image-based learning by converting the signals into scalogram images for use with advanced classification algorithms. The method described in Subsection IV.5.3, which computed scalograms for the entire signal, would not be optimal for the OSA database recordings that span an extremely long duration. Even after preprocessing to shorten the signals, their remaining length, often several hours, creates significant computational challenges and degrades the quality of the scalograms.

To preserve high-frequency details in the scalograms, we divide each signal into 1-minute segments. We then compute the CWT for each epoch using the Morlet wavelet, which offers an optimal balance between time and frequency localization. Each resulting scalogram is saved as an image, thereby creating a feature folder for each subject. Next, we consider whether to further process these images by arranging them in temporal order into a composite montage image. This montage can be organized as a two-dimensional grid, with one axis representing time progression and the other reflecting the frequency content of each epoch. The final montage can be resized and fed into a pre-trained CNN as a single image, allowing the network to capture temporal patterns over the entire recording rather than from isolated snapshots. However, if the montage is too dense, downsampling may obscure important details.

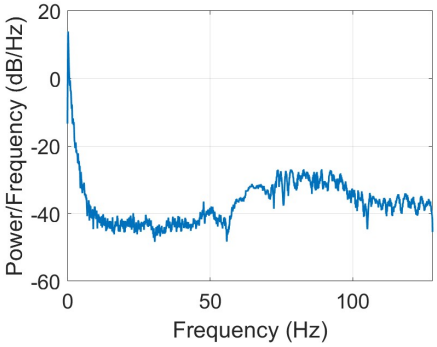
Unfortunately, the input image sizes for pre-trained ImageNet networks vary from 224 to 512 pixels, which means that the finer details of the scalograms might become obscured or even lost. A smaller input might risk losing some details, especially if respiratory events occur in narrow time windows. So we opted for an alternative method of constructing a master feature array to obtain the generalized features from the individual scalograms, which will be discussed in detail in the Chapter V and thus not requiring to create the montages of the scalograms.



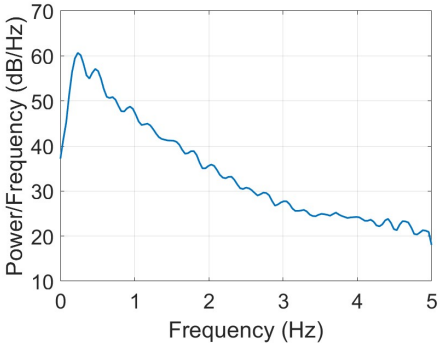

**(a)** Time-Amplitude Plot (OSAMAS_012)



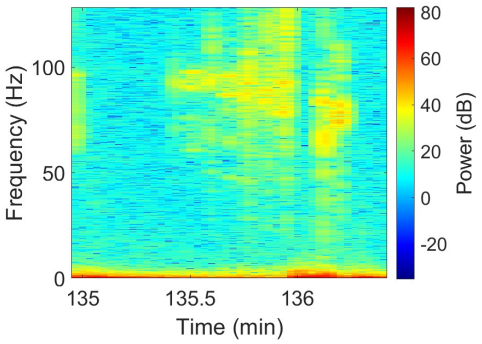

**(b)** Time-Amplitude Plot (OSAMAS_029)



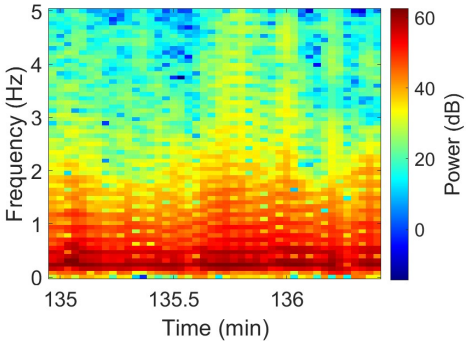

**(c)** PSD (OSAMAS_012)



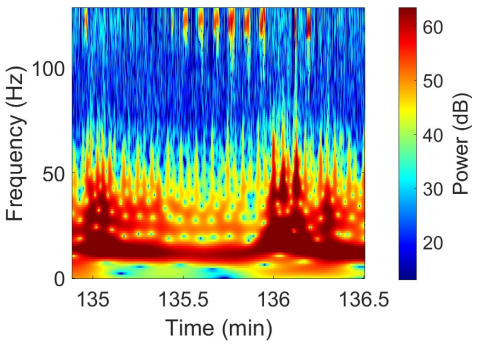

**(d)** PSD (OSAMAS_029)



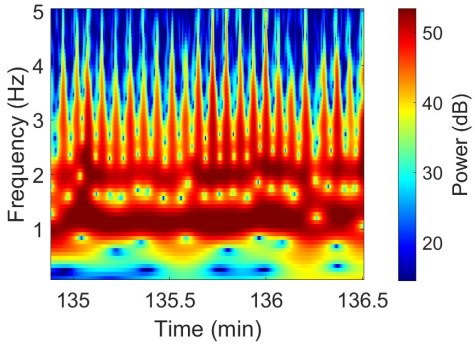

**(e)** Spectrogram (OSAMAS_012)




**(f)** Spectrogram (OSAMAS_029)




**(g)** Scalogram (OSAMAS_012)




**(h)** Scalogram (OSAMAS_029)

**Figure IV.17:** Comparison of hypopnea events between OSAMAS_012 (left column) and OSAMAS_029 (right column), showing (a–b) time-domain signals, (c–d) PSDs, (e–f) spectrograms, and (g–h) scalograms.

### IV.6.3 Features for Non-image based learning obtained from Scalograms

Before applying the sophisticated image-based learning methods to the scalogram image dataset obtained in the previous subsection, we first assessed whether there were statistically significant differences between classes within each of the OSAMAS, CRC, and PhysMAS datasets. To achieve this, texture analysis was performed as described in Subsection IV.4.3. The processing and extraction of GLCM texture features were implemented using Python (Python Core Team, 2019).

A processing pipeline was created to extract texture features from image data and to statistically compare two subject groups. In this pipeline, images belonging to each subject were first converted to grayscale before computing the GLCM. From the computed GLCM, six texture features were extracted: contrast, energy, entropy, homogeneity, correlation, and dissimilarity. These features were calculated for every image within a subject folder, and the results were subsequently aggregated by averaging to obtain subject-level metrics. The aggregated features for all subjects were compiled into a single dataset, and the resulting subject-level features were exported as `.csv` files for further statistical analysis, which will be discussed in the Chapter V.

## Chapter V

# Machine Learning Methods for Signal Classification

In this chapter, we test the hypotheses introduced in Chapter III by developing and evaluating an automatic classification pipeline for our biomedical signals. In Chapter IV, we analyzed raw signals and extracted features from skeletal muscle iEMG recordings and patient airflow measurements, thus preparing comprehensive datasets suitable for classification tasks. Here, we leverage these pre-processed and engineered features to design experiments that compare both classical machine learning and deep learning approaches, aiming to build robust and interpretable models suitable for clinical decision-making. We will define the experimental setting, select appropriate classification methods, and determine relevant evaluation metrics tailored specifically for each dataset, taking into consideration the feature types and dataset sizes.

Given that both skeletal muscle disorders datasets and the OSA dataset involve clearly labeled data, we frame our problem as a supervised classification task. Our initial objective is the development of an automated classification pipeline optimized for feature-rich biomedical signals, making machine learning algorithms particularly suited for our purposes. As such automated pipelines are intended for potential clinical implementation, we define robust experimental setups and select evaluation metrics aimed at minimizing potential risks associated with erroneous predictions.

This chapter is structured as follows:

- **Experimental Setup and Evaluation Metrics (Section V.1):** We detail cross-validation strategies, data splitting methods, and performance metrics that define our experimental framework.
- **Machine Learning Methods (Section V.2):** We present a comprehensive

overview of classifiers, covering classical algorithms and deep learning architectures.

- **Skeletal Muscle Disorders Datasets: Classification (Section V.3):** We apply our machine learning pipelines to skeletal muscle iEMG signals, evaluating feature selection strategies, augmentation techniques, and model calibration effectiveness.

- **Obstructive Sleep Apnea Dataset: Classification (Section V.4):** We will apply image-based transfer learning approaches and classical methods utilizing engineered features on OSA datasets.

## V.1 Experimental Setup

### V.1.1 Objectives

In this work, we define two distinct classification tasks: one for skeletal muscle disorders using labeled EMG data and one for OSA using polysomnography data. The methods for classification and the corresponding results are discussed in Sections V.3 and V.4, respectively. Both tasks are cast as supervised classification problems based on the labeled data obtained from our previous feature extraction procedures outlined in Chapter IV. The experimental design leverages these features and employs rigorous cross-validation techniques to ensure robust model evaluation.

For the skeletal muscle disorders datasets, our primary objective is to test a set of hypotheses developed for skeletal muscle channelopathies, refer to Subsection III.1.4. These include: (i) H1: the existence of distinct iEMG features, (ii) H2: the necessity of automated analysis, and (iii) H3: the clinical utility of novel diagnostic criteria. Feature extraction detailed on page 38, which is the Step 1 of the plan was completed in Chapter IV. In this chapter, we advance to Step 2 by developing an automated pipeline for signal classification and by evaluating the obtained results.

The classification of the fibrillation potentials dataset aims to evaluate the universal applicability of the pipeline developed for the channelopathy dataset, see the hypotheses set on page 40. In this chapter, we apply Steps 2–4, outlined on pages 40–41, to the features extracted in Chapter IV. First, we identify the most effective methods for the channelopathy dataset; then, we apply these methods to the fibrillation potentials dataset and assess their performance using appropriate evaluation metrics.

To test the hypothesis for the OSA dataset described in Subsection III.2.2, we will follow the steps outlined on page 45 of that subsection. In Chapter III and Chapter IV, we completed Step 1 by acquiring the relevant portions of the polysomnography data and Step 2 by extracting the characteristic features from patient airflow measurements, respectively. In this chapter, we proceed with Step 3, which is crucial for determining whether our hypothesis, that the shape of patient airflow is predictive of a successful

response to MAS treatment in OSA patients, holds.

The following sections present the experimental setup, the selected machine learning methods, and the evaluation metrics employed to validate our hypotheses. All code was written in Python (Python Core Team, 2019), unless otherwise specified. In particular, classical machine learning experiments were implemented using Scikit-learn (Pedregosa et al., 2011), and neural network pipelines were developed with PyTorch (Paszke et al., 2017) package.

### V.1.2 Cross-Validation, Data Splitting, and Model Assessment

The ultimate goal of any learning method is to generalize well on independent test data. In practice, we evaluate a model's generalization performance by estimating its prediction error on unseen data, which guides model selection and provides a measure of the final model's quality.

For a quantitative response, let $Y$ denote the target variable and $\mathbf{X}$ the input vector. Given a prediction model $\hat{f}(\mathbf{X})$ built using a training set $T$, performance is measured by a loss function $L(Y, \hat{f}(\mathbf{X}))$ (Hastie, 2009). Common loss functions include the squared error,

$$L(Y, \hat{f}(\mathbf{X})) = \left(Y - \hat{f}(\mathbf{X})\right)^2, \tag{V.1}$$

and the absolute error,

$$L(Y, \hat{f}(\mathbf{X})) = \left|Y - \hat{f}(\mathbf{X})\right|. \tag{V.2}$$

The *test error* or generalization error is defined as the expected loss on an independent test sample drawn from the population:

$$\mathrm{Err}_T = \mathbb{E}\left[L(Y, \hat{f}(\mathbf{X})) \mid T\right]. \tag{V.3}$$

Because both $X$ and $Y$ are random, we are often interested in the *expected prediction error*,

$$\mathrm{Err} = \mathbb{E}\left[L(Y, \hat{f}(\mathbf{X}))\right] = \mathbb{E}\left[\mathrm{Err}_T\right], \tag{V.4}$$

which averages over the randomness in both the test data and the training process.

In contrast, the training error is computed on the same data used to fit the model:

$$\mathrm{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right). \tag{V.5}$$

While training error typically decreases with model complexity, a model with very low training error might be overfitting and thus generalize poorly. Using the concept of Vapnik-Chervonenkis (VC) dimension, one can derive theoretical bounds on how much the training error underestimates the true prediction error (Hastie, 2009). The VC dimension is a measure of the complexity of a class of functions, and these bounds indicate that when fitting a model with a given VC dimension on N training points, the training error is optimistically biased: it is lower than the true error by an amount that depends on both the VC dimension and the observed error. This theory explains why a model that performs well on the training data may not generalize to new, unseen data.

Since directly estimating the true prediction error on a large, independent test set is often impractical, especially when available data is limited, cross-validation becomes a standard method for estimating prediction error (Hastie, 2009). In $K$-fold cross-validation, the dataset is divided into $K$ roughly equal parts. For each fold, the model is trained on the remaining $K-1$ parts and then evaluated on the held-out fold. The cross-validation estimate of prediction error is given by:

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}_{-\kappa(i)}(x_i)\right), \tag{V.6}$$

where $\hat{f}_{-\kappa(i)}$ denotes the model fitted without the fold containing observation $i$. When $K = N$, this approach is known as leave-one-out cross-validation. Although leave-one-out provides nearly unbiased estimates of the expected error, it can have high variance and is computationally expensive.

The choice of $K$ involves a trade-off: using a small $K$, like 5 or 10, reduces variance and computational cost, but may introduce bias if the performance of the learning method is sensitive to the training set size (Hastie, 2009). For instance, if the learning curve (which plots performance versus training set size) is steep at the available sample size, $K$-fold cross-validation might overestimate the true error. In practice, however, 5- or 10-fold cross-validation is recommended as a good compromise, as supported by both theoretical analysis and empirical studies (Breiman et al., 1992; Kohavi, 1995).

While the training error typically underestimates the true prediction error due to overfitting, cross-validation provides a more reliable estimate by averaging performance over multiple training-validation splits. This approach, combined with an understanding of model complexity, helps us select models that not only perform well on training data but are also expected to generalize effectively to new, unseen data (Hastie, 2009).

In our experiments, we employ distinct cross-validation strategies tailored to each dataset. For the skeletal muscle disorders datasets, where multiple samples per patient are available,

we use *grouped cross-validation*. This strategy ensures that all samples from the same patient appear only in either the training or the validation set, thus preventing data leakage and accurately assessing the model's ability to generalize across patients. To implement this, we developed a script that extracts patient IDs from filenames using regular expressions, constructs a directory structure for each split, and randomly assigns patients to training, validation, and test sets, while logging assignments for reproducibility. For the channelopathy dataset, we created four splits, each with roughly 50 patients in the training set and 8 patients in both the validation and test sets; similar procedures were applied to the synthetic data and fibrillation potentials datasets.

In contrast, for the OSA dataset, where only one measurement is available per patient, we adopt a *leave-one-out* approach, which is particularly suitable given the limited number of subjects and the high cost associated with misclassification. In this approach, the model is trained on all samples except one, which is then used for validation. This process is repeated until each sample in the dataset has served as the validation set exactly once. Leave-one-out maximizes the training data available in each iteration, which is particularly beneficial for small datasets. However, it can be computationally intensive and may increase the risk of overfitting if not properly regularized.

Grouped cross-validation is particularly appropriate for datasets with multiple samples per patient, such as the skeletal muscle disorders datasets, because it prevents data leakage by ensuring that all samples from a given patient are confined to either the training or the validation set, thereby providing a realistic assessment of the model's ability to generalize to unseen subjects. Conversely, leave-one-out cross-validation is ideal for the OSA dataset, where only one sample per patient is available, as it maximizes the use of the limited data. Accordingly, we adopted grouped cross-validation for the skeletal muscle disorders datasets and leave-one-out cross-validation for the OSA dataset.

### V.1.3 Evaluation Metrics

Evaluating our models requires careful consideration of class distribution, especially when working with imbalanced datasets. For example, our channelopathy dataset comprises two classes – sodium and chloride, with 279 and 221 samples respectively, resulting in a mild imbalance ratio of approximately 1.26:1. Similarly, the OSA datasets exhibit an imbalance, with non-responder samples predominating. The combined OSA dataset, including all three subsets, comprises 47 responders and 55 non-responders (imbalance ratio of 1.17:1), whereas the CRC and PhysMAS subsets alone have 33 responders and 45 non-responders (imbalance ratio of 1.36:1). Given these conditions, maintaining consistent class distributions across cross-validation splits is essential to achieve reliable performance estimates.

While accuracy – the proportion of correctly classified samples, provides a general measure

of performance, it can be misleading in imbalanced scenarios where a model may achieve high accuracy simply by favoring the majority class. To address this limitation, we complement accuracy with additional metrics that are more sensitive to class imbalances. In our evaluation, we compute metrics such as precision, recall, F1-score, and ROC-AUC, along with others specifically tailored for imbalanced data. All metrics are derived from cross-validation validation sets to ensure an unbiased assessment of generalization performance.

Precision measures the proportion of true positive predictions among all positive predictions, reflecting the model's ability to minimize false positives. In contrast, recall (or sensitivity) quantifies the proportion of actual positive instances that are correctly identified, which is essential for detecting the minority class. The F1-score – the harmonic mean of precision and recall, provides a balanced performance measure, especially valuable when class distributions are skewed. Specificity, defined as the proportion of true negatives correctly identified, indicates how likely the model is to classify a truly negative instance as negative. This metric complements sensitivity by evaluating performance on both classes. High specificity is particularly crucial in biomedical contexts, where misclassifying a normal signal as pathological could lead to unnecessary follow-up. Although specificity can be deceptively high in imbalanced datasets dominated by negatives, examining it alongside sensitivity offers a more balanced and informative view of classifier performance.

Balanced accuracy provides a comprehensive summary of classifier performance by accounting for class imbalances. It is defined as the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), thereby giving equal importance to both classes. This metric is especially critical in scenarios where the majority class dominates; for example, in a dataset with 98 % negatives, a model predicting only the majority class might still achieve an overall accuracy of approximately 98 % while failing to detect any minority instances. In contrast, balanced accuracy would reveal this shortfall by averaging a low sensitivity with a high specificity, potentially resulting in a score around 60 %. In our cross-validation analysis, we rely on balanced accuracy to ensure that the model demonstrates robust performance across both normal and abnormal signal classes.

An additional metric commonly used for classifier evaluation is the Area Under the Receiver Operating Characteristic (ROC) Curve (ROC-AUC) (Fawcett, 2006). In binary classification, each instance is assigned a positive or negative label, and classifiers produce predictions either as continuous probability scores, requiring a threshold to yield discrete outcomes, or as direct class labels. Performance is typically summarized in a confusion matrix, which tabulates true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). From this matrix, the true positive rate (TPR, also known as

recall) and the false positive rate (FPR) are computed as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{V.7}$$

The ROC curve plots TPR against FPR for various threshold settings, with the ROC-AUC providing an overall measure of the classifier's ability to distinguish between classes. However, in imbalanced settings this metric can be overly optimistic.

To address class imbalance more effectively, we consider the precision-recall curve, which plots precision against recall and focuses on the classifier's performance for the positive class. This metric is generally more informative than the ROC curve in imbalanced settings, as it directly reflects the model's ability to identify the minority class. The area under the precision-recall curve, known as average precision, summarizes the trade-off between precision and recall across varying thresholds. It is especially meaningful for imbalanced datasets because its baseline is determined by the prevalence of the positive class, for example if 12 % of samples are positive, then a random classifier would yield an average precision of approximately 0.12, whereas the baseline for ROC-AUC is 0.5 for a random classifier. Thus, an average precision value significantly higher than the positive class prevalence indicates effective detection of the minority class. Reporting both average precision and ROC-AUC provides a comprehensive evaluation of classifier performance, particularly in identifying positive instances within imbalanced data.

Cohen's kappa, denoted as $\kappa$, is a statistic that measures the agreement between predicted and true labels while accounting for the agreement expected by chance. Originally developed for inter-rater reliability, this metric has become valuable for classifier evaluation, particularly in imbalanced scenarios where the potential for "chance agreement" (such as always predicting the majority class) is high. The kappa statistic is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{V.8}$$

where $p_o$ is the observed accuracy and $p_e$ is the expected accuracy by chance (calculated based on class frequencies). By subtracting $p_e$, $\kappa$ penalizes models that merely mirror the prevalent class distribution. A $\kappa$ value of 1 indicates perfect agreement beyond chance, while $\kappa = 0$ suggests that the classifier performs no better than random guessing given the class imbalance. In some cases, $\kappa$ can even be negative if the model performs worse than chance.

In our context, Cohen's kappa provides a more strict evaluation than simple accuracy. For example, a classifier achieving 90 % accuracy in a dataset where 90 % of the signals are normal might only yield a $\kappa$ of around 0.5 or lower, highlighting that much of its accuracy stems from consistently predicting the majority class. Conversely, a high $\kappa$ value

indicates that the model is capturing meaningful signal beyond what a frequency-based guess would achieve. Therefore, we include Cohen's kappa in our evaluation to ensure that the classifier's performance reflects genuine predictive power across classes, rather than merely an artifact of class imbalance.

We further assess model confidence by examining the entropy of its predicted probability distribution. For probabilistic classifiers, such as neural networks with a softmax output, the output probabilities indicate the model's level of certainty. We quantify this uncertainty using Shannon entropy,

$$H(p) = -\sum_{c} p(c) \log p(c), \tag{V.9}$$

where a low entropy value signifies that the probability mass is concentrated on a single class (indicating high confidence), and a high entropy value reflects a more uniform distribution (indicating uncertainty). In our evaluation, we analyze both the average predictive entropy and its distribution across the validation set to gauge how the model's confidence correlates with its performance.

This entropy measure provides insights into the model's internal confidence independent of prediction correctness. For instance, a well-calibrated model might exhibit high entropy for ambiguous or difficult cases and low entropy for clear-cut examples where evidence strongly favors one class. When comparing models with similar accuracy, the one with consistently lower entropy can be considered to make more decisive predictions. However, entropy should not be used in isolation; it is most informative when considered alongside accuracy and other performance metrics, ensuring that the classifier is both precise and well-calibrated in its confidence.

By combining these metrics, we achieve a comprehensive assessment of classifier performance under class imbalance. Specificity and sensitivity show how each class is handled, while balanced accuracy and Cohen's kappa offer chance-corrected summaries that account for class frequency disparities. Furthermore, average precision directly measures the model's ability to identify the minority class, and entropy-based confidence analysis reveals whether the model's probabilistic predictions are well-calibrated. Together, these metrics, computed on cross-validation validation sets, provide a robust framework for evaluating model effectiveness and reliability in our imbalanced biomedical signal classification task. This multifaceted evaluation is critical in high-stakes clinical settings, ensuring that the chosen model not only achieves high overall accuracy but also effectively detects the minority class and demonstrates appropriate confidence in its predictions.

**Unless otherwise specified, rate-type metrics (accuracy, balanced accuracy, precision, recall, specificity, and F1 score) are reported as percentages, whereas**

**probabilistic or agreement metrics (Cohen's kappa, ROC-AUC, predictive confidence, and entropy) are given in their native 0-1 range.**

## V.2 Machine Learning Methods

In this section, we provide a comprehensive overview of the machine learning methods used to classify our biomedical signal datasets. We frame our problem within the supervised learning paradigm, where a defined outcome variable guides the learning process (Hastie, 2009). Specifically, our task is formulated as a binary classification problem, as both the skeletal muscle disorder datasets (excluding the simulated dataset) and the OSA datasets involve distinguishing between two classes.

We begin by outlining the spectrum of methods that form the backbone of predictive modeling. Classical approaches such as logistic regression, support vector machines, random forests, and eXtreme gradient boosting are introduced first. These methods are especially well-suited for scenarios in which the data are structured and tabular, and where interpretability is a key concern. Their solid theoretical foundations and robustness make them ideal when the dataset is limited or when the relationships among variables are relatively straightforward.

Following the discussion of classical methods, the focus shifts to neural network architectures, which have emerged as powerful tools in settings where data exhibit complex, hierarchical structures. Neural networks, including convolutional neural networks and transformer-based models, excel in extracting and combining features from raw data, such as images and time-series scalograms. Their ability to learn multi-level representations enables state-of-the-art performance in diverse biomedical applications, despite their often higher computational demands and larger data requirements.

Complementing these approaches, we also introduce advanced techniques such as transfer learning and ensemble methods, which play a crucial role in enhancing model robustness and generalization—especially when training data are scarce. Data augmentation strategies further bolster this effort by expanding the effective training set with realistic variations, thereby mitigating overfitting and supporting the development of more resilient predictive models.

This progression from classical machine learning methods to advanced neural network architectures sets the stage for exploring even more specialized deep learning models and hybrid approaches in the next section.

### V.2.1 Classical Machine Learning Methods

Classical machine learning methods have long formed the foundation of predictive modeling by offering robust and interpretable approaches to classification and regression.

These methods, grounded in statistical theory, are particularly effective when applied to structured, tabular data with moderate feature spaces, encompassing continuous, categorical, or ordinal variables. Being an alternative to a deep learning, they are based on clear model assumptions that both facilitate accurate prediction and enhance our understanding of data relationships.

A key element in understanding classical methods is recognizing how they are classified based on their learning paradigms and underlying model assumptions. For example, logistic regression operates within a supervised learning framework as a parametric model because it directly estimates class probabilities using a fixed functional form. In contrast, support vector machines are considered non-parametric since they adapt to the data without a predetermined structure by maximizing the margin between classes to define the decision boundary.

Ensemble methods provide another dimension to this classification. Random forests, for example, aggregate multiple decision trees to form a non-parametric model that adapts flexibly to the data, effectively capturing complex, non-linear relationships. Similarly, eXtreme gradient boosting employs a sequential boosting strategy, iteratively combining decision trees to reduce prediction errors and enhance overall performance.

These classical methods provide a balanced spectrum of tools that prioritize interpretability, robustness, and predictive accuracy. Their application spans diverse domains, including medical diagnosis and bioinformatics, where understanding the influence of individual predictors can be critical. In the following sections, we will explore logistic regression, support vector machines, random forests, and eXtreme gradient boosting in greater detail, discussing their underlying principles, optimization strategies, and practical considerations, with reference to (Hastie, 2009).

## Logistic Regression

Logistic regression is a linear method used for classification that builds a probabilistic model for predicting the likelihood that a given input belongs to the positive class. Specifically, it models the conditional probability $p(y = 1 \mid \mathbf{x})$ as a sigmoid transformation of a linear combination of the input features. Given a feature vector $\mathbf{x} \in \mathbb{R}^d$ and a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$, which includes an intercept term $\beta_0$, the model is defined as

$$p(y = 1 \mid \mathbf{x}; \boldsymbol{\beta}) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_j\right)\right)}. \tag{V.10}$$

This formulation with sigmoid function ensures that the predicted probability lies between 0 and 1 (Hastie, 2009).

The decision boundary of the model is a hyperplane and is defined by

$$\beta_0 + \sum_{j=1}^{d} \beta_j x_j = 0. \tag{V.11}$$

A key advantage of logistic regression is its interpretability. The model not only provides probability estimates for each class, but its coefficients can be exponentiated to yield odds ratios. Specifically, $e^{\beta_j}$ quantifies the multiplicative change in the odds of the outcome associated with a one-unit increase in the $j$th predictor, making it easier to understand the influence of individual features on the response.

The logistic regression model is fitted by maximizing the conditional log-likelihood of the observed data. For a binary classification problem, where each response $y_i$ is either 0 or 1, the log-likelihood function is defined as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[ y_i \log p(\mathbf{x}_i; \boldsymbol{\beta}) + (1 - y_i) \log \left(1 - p(\mathbf{x}_i; \boldsymbol{\beta})\right) \right]. \tag{V.12}$$

To find the parameter vector $\boldsymbol{\beta}$ that maximizes $\ell(\boldsymbol{\beta})$, we set the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ equal to zero. This yields the score equations:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - p(\mathbf{x}_i; \boldsymbol{\beta})\right) = 0. \tag{V.13}$$

Because these equations are nonlinear in $\boldsymbol{\beta}$, we employ an iterative method such as the Newton-Raphson algorithm to obtain the maximum likelihood estimates.

In matrix notation, let $\mathbf{y}$ be the $N \times 1$ vector of responses, $\mathbf{X}$ be the $N \times (d+1)$ design matrix, which includes a column of ones for the intercept, and $\mathbf{p}$ be the $N \times 1$ vector of fitted probabilities. The gradient of the log-likelihood can then be written as

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^\top (\mathbf{y} - \mathbf{p}), \tag{V.14}$$

and the Hessian (second derivative) is given by

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\mathbf{X}^\top \mathbf{W} \mathbf{X}, \tag{V.15}$$

where $\mathbf{W}$ is a diagonal matrix with entries $W_{ii} = p(\mathbf{x}_i; \boldsymbol{\beta}) \left(1 - p(\mathbf{x}_i; \boldsymbol{\beta})\right)$.

Using the Newton-Raphson method, the parameter vector is updated iteratively according

to

$$\boldsymbol{\beta}_{\text{new}} = \boldsymbol{\beta}_{\text{old}} - \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p}). \tag{V.16}$$

This update rule forms the basis of the iteratively reweighted least squares (IRLS) algorithm, which is repeated until convergence is achieved (Hastie, 2009). Each iteration refines the estimate of $\boldsymbol{\beta}$ by re-calculating the probabilities $\mathbf{p}$ and the weight matrix $\mathbf{W}$, ensuring that the parameter estimates converge to the maximum likelihood solution.

Logistic regression is widely used both for prediction and for inferential analysis, as it provides interpretable measures that help understand the influence of predictors on the outcome. However, its validity rests on several key assumptions. First, the model assumes that the log-odds of the outcome are a linear function of the predictors; if this assumption is violated, the model may misrepresent the true relationship between predictors and the outcome. Second, the observations must be independent – any correlation, such as that introduced by repeated measures or clustered data, can lead to biased estimates and erroneous inference. Third, there must be no perfect separation between classes, as perfect separation undermines the stability and existence of the maximum likelihood estimates. When these assumptions are violated or when the data are high-dimensional with complex interactions, alternative methods, such as regularized logistic regression or ensemble techniques, may yield improved performance.

### Support Vector Machines

Support vector machines (SVM) provide a geometrically motivated approach to binary classification by focusing on the optimal placement of the decision boundary. Unlike logistic regression, which directly models class probabilities using a sigmoid function, SVM aim to find a hyperplane that maximizes the *margin* – the shortest distance between the hyperplane and any training sample (vector).

Consider a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where each input $\mathbf{x}_i \in \mathbb{R}^p$ and the corresponding label $y_i \in \{-1, 1\}$. A hyperplane in the input space is defined by

$$f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0, \tag{V.17}$$

where $\boldsymbol{\beta}$ is the coefficient vector and $\beta_0$ is the intercept.

For perfectly separable data, we can choose $\boldsymbol{\beta}$ and $\beta_0$ such that

$$y_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0\right) \geq 1, \quad \forall i. \tag{V.18}$$

This normalization fixes the scale of $\boldsymbol{\beta}$ and $\beta_0$ so that the distance from any point $\mathbf{x}_i$ to the hyperplane (V.17) is given by distance $= \frac{\left|\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0\right|}{\|\boldsymbol{\beta}\|}$. Under the constraint in (V.18),

the margin is $M = \frac{1}{\|\boldsymbol{\beta}\|}$. Maximizing this margin or equivalently, minimizing $\|\boldsymbol{\beta}\|$, results in a robust classifier that is less sensitive to small perturbations in the data.

In many practical applications, however, the classes overlap, and perfect linear separation is not possible. To address this, SVM introduce nonnegative slack variables $\xi_i \geq 0$ that allow some training points to lie within or even on the wrong side of the ideal margin. The classification constraint is then relaxed to

$$y_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0\right) \geq 1 - \xi_i, \quad \forall i. \tag{V.19}$$

In this formulation, if a point violates the margin requirement, the corresponding $\xi_i$ quantifies the degree of violation.

The soft-margin SVM seeks to balance maximizing the margin and minimizing the total slack by solving the following optimization problem:

$$\min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{subject to} \quad \begin{cases} y_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0\right) \geq 1 - \xi_i, & i = 1, \ldots, N, \\ \xi_i \geq 0, & i = 1, \ldots, N. \end{cases} \tag{V.20}$$

Here, the term $\frac{1}{2}\|\boldsymbol{\beta}\|^2$ is minimized to maximize the margin, while the parameter $C > 0$ controls the trade-off between achieving a wide margin and penalizing margin violations.

The solution to the above convex optimization problem can be expressed in dual form, where the weight vector is represented solely in terms of the training data. Specifically, one can show that

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i \, \mathbf{x}_i, \tag{V.21}$$

where the Lagrange multipliers $\hat{\alpha}_i$ are nonzero only for the *support vectors* – the training points that lie on or inside the margin. Consequently, the decision function is

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i \, \langle \mathbf{x}_i, \mathbf{x} \rangle + \hat{\beta}_0, \tag{V.22}$$

and the predicted class label is obtained by

$$\hat{G}(\mathbf{x}) = \operatorname{sign}\left[\hat{f}(\mathbf{x})\right]. \tag{V.23}$$

To handle nonlinear decision boundaries, SVM employ the kernel trick. Instead of operating directly in the input space, the data is implicitly mapped into a high-dimensional feature

space via a transformation $\phi(\mathbf{x})$. In this new space, the classes may become linearly separable even if it is not possible in the original space. Rather than explicitly computing $\phi(\mathbf{x})$ for each input, which could be computationally expensive, we compute the inner product between transformed points using a kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle. \tag{V.24}$$

This kernel function measures the similarity between points in the feature space. The SVM decision function can then be expressed in terms of these kernel evaluations:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i \, K(\mathbf{x}, \mathbf{x}_i) + \hat{\beta}_0, \tag{V.25}$$

Common kernel functions include the linear kernel $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$, the radial basis function kernel $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{z}\|^2\right)$, and the polynomial kernel $K(\mathbf{x}, \mathbf{z}) = (\gamma \, \mathbf{x}^\top \mathbf{z} + r)^d$.

The geometric interpretation of SVM is central to their effectiveness. By seeking to maximize the margin, SVM not only identify a hyperplane that separates the classes but also achieve robustness to small variations in the data. The support vectors, which are the points closest to the hyperplane, play a crucial role because they uniquely determine the position and orientation of the decision boundary. This focus on maximizing the geometric margin serves as a form of regularization, helping to reduce the risk of overfitting in moderate-dimensional spaces (Cortes et al., 1995).

In contrast to logistic regression, which directly estimates class probabilities via a sigmoid function, SVM emphasize the underlying structure of the data. This makes them particularly effective in high-dimensional settings or in cases where there is a clear separation between classes. However, SVM can become computationally expensive for very large datasets due to their quadratic or cubic time complexity with respect to the number of samples, and their performance is sensitive to the choice of parameters such as the regularization constant and kernel parameters.

### Random Forests

Random forests (Breiman, 2001) build on the fundamental idea of tree-based methods, which partition the feature space into a set of regions and then fit a simple model, often a constant, in each region. In a decision tree, the algorithm recursively divides the input space into rectangular regions. The quality of each split is measured by a node impurity criterion. For classification tasks, two common measures are the *Gini index* and *cross-entropy* or deviance. The Gini index is defined as

$$G = 1 - \sum_{c=1}^{C} p_c^2, \tag{V.26}$$

where $p_c$ is the proportion of samples in the node that belong to class $c$. It can be interpreted as the expected error rate if an observation were randomly classified according to the class distribution in the node. Alternatively, the cross-entropy is given by

$$-\sum_{c=1}^{C} p_c \log(p_c). \tag{V.27}$$

Both of these measures are differentiable and sensitive to changes in class probabilities, which typically leads to more informative splits.

A critical aspect of building a decision tree is controlling its size or depth. A very large tree may overfit the data by capturing noise, whereas a tree that is too small might miss important structure. To address this, one common strategy is to grow a large tree, until a minimum node size is reached, and then prune it using cost-complexity criteria. However, single decision trees are known to be high-variance estimators; the hierarchical nature of the splitting process means that small variations in the training data can lead to substantially different trees.

Random forests address the inherent instability of individual decision trees by aggregating the predictions of many trees, each built from a different bootstrap sample of the original data. In addition to bagging, random forests introduce randomness by selecting, at each split, only a random subset of the available features. Typically, if there are $d$ features, a subset of roughly $\sqrt{d}$ features is considered. This dual mechanism – bootstrap aggregation and random feature selection, reduces the correlation among individual trees, thereby lowering the overall variance of the ensemble without a significant increase in bias. By combining these trees, random forests harness the flexible, nonparametric modeling capability of decision trees while benefiting from the stabilizing effect of ensemble learning. In classification tasks, the final prediction for an input $\mathbf{x}$ is determined by taking the majority vote across all trees in the ensemble.

Random forests are especially well-suited for high-dimensional data and problems with complex feature interactions. They naturally handle mixed data types (continuous, categorical, ordinal) and provide insights into feature importance via methods such as permutation testing. However, despite their strong predictive performance, they can be less interpretable than simpler models. Moreover, the ensemble nature of random forests can result in large model sizes, which may limit their applicability in real-time or resource-constrained environments.

### eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) (Chen et al., 2016) is an advanced ensemble learning algorithm that belongs to the family of gradient boosting methods. The core idea behind boosting is to combine many “weak” classifiers, each performing only slightly

better than random guessing, into a powerful committee (Hastie, 2009). In boosting, a weak classifier is applied sequentially to modified versions of the data, and its prediction is weighted by a coefficient that reflects its accuracy. The final prediction is then obtained by a weighted majority vote:

$$\hat{G}(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m \, G_m(\mathbf{x})\right), \tag{V.28}$$

where $G_m(\mathbf{x})$ denotes the $m$th weak classifier, $\alpha_m$ is its corresponding weight, and $M$ is the total number of boosting iterations. At each iteration, the training observations are reweighted so that misclassified examples gain more influence in subsequent rounds, thereby forcing later classifiers to focus on harder-to-classify cases.

XGBoost extends these boosting principles by using decision trees as the weak learners in a sequential framework. Unlike ensemble methods such as Random Forests that build trees in parallel, XGBoost constructs trees one after the other; each new tree $f_m$ is built to correct the residual errors of the current ensemble, which helps reduce bias while controlling model complexity.

The algorithm requires the minimization of an objective function that combines a loss term with a regularization term. Formally, the objective function is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} l\big(y_i, \hat{y}_i\big) + \sum_{m=1}^{M} \Omega\big(f_m\big), \tag{V.29}$$

where $l(y_i, \hat{y}_i)$ is a differentiable loss function that measures the discrepancy between the true label $y_i$ and the prediction $\hat{y}_i$; $f_m$ denotes the $m$th decision tree – weak learner; and $\Omega(f_m)$ is a regularization term that penalizes the complexity of $f_m$, such as the number of leaves or the magnitude of leaf weights, and $N$ is the number of training instances.

To add a new tree at iteration $m$, XGBoost uses a second-order Taylor series expansion to approximate the change in the loss function. Let $\hat{y}_i^{(m-1)}$ be the prediction from the ensemble up to iteration $m-1$. Then, the objective at iteration $m$ is approximated as

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{N} \left[ g_i \, f_m(\mathbf{x}_i) + \frac{1}{2} \, h_i \, f_m^2(\mathbf{x}_i) \right] + \Omega\big(f_m\big), \tag{V.30}$$

where the first- and second-order derivatives are defined as

$$g_i = \frac{\partial l\big(y_i, \hat{y}_i^{(m-1)}\big)}{\partial \hat{y}_i^{(m-1)}}, \quad h_i = \frac{\partial^2 l\big(y_i, \hat{y}_i^{(m-1)}\big)}{\partial \left(\hat{y}_i^{(m-1)}\right)^2}.$$

These derivatives, calculated based on the predictions of the ensemble at the previous iteration, allow for more precise updates than traditional gradient boosting methods that rely solely on first-order information.

The method incorporates several modifications to enhance performance. Its regularized learning framework prevents overfitting by penalizing overly complex trees, and its algorithm is designed to handle missing values efficiently through sparsity-aware split finding. Furthermore, it uses advanced techniques such as weighted quantile sketching to manage large datasets effectively.

While XGBoost typically requires more computational resources than simpler models like logistic regression or ensemble methods such as Random Forests, its scalability can be improved through distributed computing and GPU acceleration. This makes it particularly suitable for large-scale problems involving complex, nonlinear interactions, such as genomic data analysis or time-series classification of physiological signals.

XGBoost represents a significant evolution in boosting algorithms by sequentially building decision trees that minimize both prediction error and model complexity. By leveraging second-order optimization and regularization, it offers robust, nonlinear modeling capabilities that complement the strengths of other classical methods. In the next section, we will transition to neural networks, which take a fundamentally different approach to capturing complex patterns in high-dimensional data.

### V.2.2 Neural Networks

Neural networks represent a paradigm shift from classical machine learning methods: they automatically learn hierarchical feature representations directly from raw data through layered nonlinear transformations. Originally inspired by the structure of the human brain, where neurons communicate via synapses, these models mimic the process of information processing by using units (neurons) and weighted connections (synapses). In essence, neural networks learn to extract and combine features to capture complex patterns and relationships. The following description of the principle algorithms is given according to (Hastie, 2009).

The core idea is to transform the original input features into a set of higher-level, derived features using learned linear combinations, and then model the target variable as a nonlinear function of these derived features. In a typical single hidden layer network, the process is divided into two stages.

In the first stage, for each hidden unit $m$ (with $m = 1, \ldots, M$), the network computes a pre-activation value and then applies a nonlinear activation function. Specifically, given

an input vector $\mathbf{x} \in \mathbb{R}^d$, the pre-activation is computed as

$$z_m = \alpha_{0m} + \boldsymbol{\alpha}_m^\top \mathbf{x}, \tag{V.31}$$

and the activated output is

$$a_m = \sigma(z_m), \tag{V.32}$$

where $\sigma(\cdot)$ is a nonlinear activation function (commonly the sigmoid $\sigma(v) = \frac{1}{1+e^{-v}}$), $\alpha_{0m}$ is the bias, and $\boldsymbol{\alpha}_m \in \mathbb{R}^d$ is the corresponding weight vector. The outputs of all hidden units are collected into the feature vector $\mathbf{a} = (a_1, a_2, \ldots, a_M)^\top$.

In the second stage, these features are linearly combined to produce the network output. For regression (or as an intermediate representation in classification), the output is given by

$$t = \beta_0 + \boldsymbol{\beta}^\top \mathbf{a}, \tag{V.33}$$

where $\boldsymbol{\beta} \in \mathbb{R}^M$ is the weight vector connecting the hidden layer to the output and $\beta_0$ is the output bias. For classification with $K$ classes, the network typically employs $K$ output units with

$$t_k = \beta_{0k} + \boldsymbol{\beta}_k^\top \mathbf{a}, \quad k = 1, \ldots, K, \tag{V.34}$$

and a softmax function converts these raw outputs into class probabilities:

$$P(Y = k \mid \mathbf{x}) = \frac{\exp(t_k)}{\sum_{\ell=1}^{K} \exp(t_\ell)}. \tag{V.35}$$

The complete set of network parameters is denoted by

$$\theta = \Big\{\alpha_{0m},\, \boldsymbol{\alpha}_m \,:\, m = 1, \ldots, M\Big\} \cup \Big\{\beta_{0k},\, \boldsymbol{\beta}_k \,:\, k = 1, \ldots, K\Big\}. \tag{V.36}$$

Training the network involves minimizing a *loss function* $\mathcal{L}(\theta)$ that quantifies the discrepancy between the true outputs and the network's predictions. For instance, in a $K$-class classification problem, the cross-entropy loss is often used:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P(Y = k \mid \mathbf{x}_i), \tag{V.37}$$

where $y_{ik}$ is a binary indicator that equals 1 if observation $i$ belongs to class $k$, and 0 otherwise.

The backpropagation algorithm computes the gradient of $\mathcal{L}(\theta)$ with respect to each

parameter by applying the chain rule in a layer-wise fashion. For a general multilayer network, denote the input layer by $\mathbf{a}^{(0)} \equiv \mathbf{x}$ and let $\mathbf{a}^{(l-1)}$ be the activations from the previous layer. The pre-activation vector in layer $l$ is computed as

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \tag{V.38}$$

and the activations are obtained via

$$\mathbf{a}^{(l)} = \sigma\big(\mathbf{z}^{(l)}\big). \tag{V.39}$$

Here, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ denote the weight matrix and bias vector for layer $l$, respectively.

Defining the error term for layer $l$ as

$$\delta^{(l)} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{(l)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(l)}} \odot \sigma'\big(\mathbf{z}^{(l)}\big), \tag{V.40}$$

the gradient with respect to the weight matrix is then given by

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} = \delta^{(l)}\mathbf{a}^{(l-1)\top}. \tag{V.41}$$

The backpropagation algorithm consists of a forward pass, in which the network computes the activations for each layer to produce the final output, and a backward pass, where the error at the output is propagated backward through the network using (V.41) to compute the gradients. These gradients are then used to update the network parameters via an optimization algorithm, such as stochastic gradient descent, over multiple training epochs until convergence.

This hierarchical, nonlinear structure enables neural networks to approximate complex functions – a property guaranteed by the universal approximation theorem, making them powerful tools for a wide range of tasks, from image classification and speech recognition to biomedical data analysis. However, training deep networks can be computationally intensive and may require techniques such as regularization, dropout, or early stopping to prevent overfitting.

### Convolutional Neural Networks

Building upon the foundation of dense neural networks, Convolutional Neural Networks (CNN) are designed to process grid-structured data like images. In contrast to fully connected architectures where each neuron interacts with all inputs, CNN exploit the spatial structure of the data by enforcing local connectivity and parameter sharing. This design not only reduces the number of trainable parameters but, when combined with pooling operations, helps to achieve a degree of translational invariance.

A typical CNN is composed of multiple stages, each of which transforms an input set of arrays, commonly referred to as *feature maps*, into a new set of feature maps. For example, when processing a color image, the input can be viewed as a three-dimensional tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $n_1$ denotes the number of channels and $n_2 \times n_3$ represents the spatial dimensions. Each stage is generally built from three types of layers: a filter bank layer, a non-linearity layer, and a feature pooling layer (LeCun et al., 2010).

In the filter bank layer, the network applies a set of learnable convolutional filters or kernels to the input feature maps. Let $\mathbf{X}_i$ denote the $i$th input feature map and consider a kernel $\mathbf{K}_{ij} \in \mathbb{R}^{l_1 \times l_2}$ that connects the $i$th input to the $j$th output feature map. The convolutional operation computes the $j$th output feature map $\mathbf{Y}_j$ as

$$\mathbf{Y}_j(p,q) = b_j + \sum_{i=1}^{n_1} \sum_{m=1}^{l_1} \sum_{n=1}^{l_2} \mathbf{K}_{ij}(m,n)\, \mathbf{X}_i(p+m-1,\, q+n-1), \tag{V.42}$$

where $b_j$ is a trainable bias parameter and the indices $p$ and $q$ traverse the spatial dimensions. This formulation guarantees that the same filter is applied at every spatial location, thereby imparting translation equivariance to the convolution operation.

Following convolution, a pointwise non-linearity is applied to the output feature maps. While early implementations of CNN employed activation functions such as the sigmoid or hyperbolic tangent, modern architectures often favor the rectified linear unit (ReLU) or its variants, which help mitigate the vanishing gradient problem and promote sparse activations. In some cases, the non-linearity is accompanied by local normalization operations, such as subtractive and divisive normalization, to enhance contrast and encourage competition among neighboring features (LeCun et al., 2010).

Subsequent to the non-linearity, a pooling layer is typically introduced to reduce the spatial resolution of the feature maps and to provide a degree of translational invariance. A common approach is max-pooling, which is defined as

$$\mathbf{P}_j(p,q) = \max_{(u,v) \in \mathcal{N}(p,q)} \mathbf{H}_j(u,v), \tag{V.43}$$

where $\mathbf{H}_j$ represents the activation of the $j$th feature map following the non-linearity, and $\mathcal{N}(p,q)$ denotes a local neighborhood around the spatial position $(p,q)$. Alternatively, average pooling may be employed to compute the mean value within each neighborhood (LeCun et al., 2010). Both strategies serve to decrease computational complexity in deeper layers while maintaining the robustness of the learned features to small spatial variations.

By stacking multiple convolutional stages, CNN progressively build a hierarchy of features. Early layers tend to capture simple patterns such as edges and textures, whereas deeper layers encode more complex structures and object parts. The final stages of the network

typically include one or more fully connected layers that integrate the spatially distributed features to perform classification or regression tasks.

Training CNN follows the same principles as for dense neural networks. The entire set of parameters, including the convolutional kernels, biases, and any parameters associated with normalization or pooling, is optimized using stochastic gradient descent combined with backpropagation. In doing so, the network automatically learns to extract and combine features that are most relevant to the task at hand.

The key features of CNN lie in their use of local receptive fields, weight sharing, and pooling operations. These elements not only lead to significant reductions in the number of parameters but also enable the network to effectively model the spatial structure inherent in grid-like data, which has been instrumental in succeeding in image recognition.

### Vision Transformer Models

Vision Transformers (ViT) extend the Transformer architecture (Vaswani, 2017), originally designed for natural language processing, to computer vision. Unlike CNN, which encode strong inductive biases such as translation equivariance and locality, ViT treat an image as a sequence of patches and learn spatial relationships directly from data using self-attention (Dosovitskiy et al., 2020).

Traditional Transformer models follow an encoder-decoder structure. In such models, the encoder first converts an input sequence of symbol embeddings into a sequence of continuous representations. Then, the decoder generates an output sequence one element at a time in an autoregressive manner (Graves, 2013), where each element is generated based on the previously generated ones.

The key operation is the self-attention mechanism, which allows the model to weigh the importance of different input elements when constructing each output representation. In particular, for each input token, the model computes three vectors: a *query* vector, which represents the token's request for information, a *key* vector, which represents the token's content, and a *value* vector, which contains the information to be aggregated. These vectors are arranged in matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. The self-attention output is computed as a weighted sum of the values, where the weights are determined by the scaled dot products between the query and key vectors:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\Big)\mathbf{V}, \tag{V.44}$$

where $d_k$ is the dimensionality of the key vectors (Vaswani, 2017).

The Transformer enhances this mechanism through *multi-head attention*. Instead of

computing a single attention function, the model projects the inputs into multiple subspaces and computes attention in parallel: each projection is called a *head*. The outputs from all heads are then concatenated and projected back to the desired dimension:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O, \tag{V.45}$$

with each head computed as

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \tag{V.46}$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are learnable projection matrices for the $i$th head, and $\mathbf{W}^O$ is the output projection matrix (Vaswani, 2017).

To incorporate positional information, since the Transformer has no built-in notion of order, positional embeddings are added to the token embeddings. These embeddings provide the model with information about the relative or absolute position of each token in the sequence.

To adapt Transformers for vision tasks (Dosovitskiy et al., 2020), an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is first divided into a sequence of flattened patches. Suppose each patch has dimensions $P \times P$; then the total number of patches is $N = \frac{HW}{P^2}$ and each patch is reshaped into a vector in $\mathbb{R}^{P^2\cdot C}$. These vectors are projected into a latent space of dimension $D$ using a trainable linear projection:

$$\mathbf{z}_p = \mathbf{E}\,\text{vec}(\mathbf{x}_p), \quad \mathbf{E} \in \mathbb{R}^{D\times(P^2\cdot C)}, \tag{V.47}$$

where $\text{vec}(\mathbf{x}_p)$ denotes the vectorized patch. A learnable classification token $\mathbf{z}_0$ is then prepended to the sequence, and learnable position embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ are added to retain spatial information:

$$\mathbf{Z}^0 = \begin{bmatrix}\mathbf{z}_0 & \mathbf{z}_1 & \cdots & \mathbf{z}_N\end{bmatrix}^{\mathrm{T}} + \mathbf{E}_{\text{pos}}. \tag{V.48}$$

This sequence $\mathbf{Z}^0$ serves as the input to the Transformer encoder, which consists of $L$ layers of multi-head self-attention and multi-layer perceptron (MLP) blocks. Each encoder layer typically applies layer normalization before the attention and MLP blocks and includes residual connections after each block. The self-attention mechanism allows the model to capture long-range dependencies across the entire image by computing relationships between all pairs of patches.

For classification, the output corresponding to the classification token, $\mathbf{z}_0^L$, is used as a global image representation. This representation is then passed through a classification head to produce the final prediction. During pre-training, large ViT models are trained on extensive datasets, and for downstream tasks, the pre-trained model is fine-tuned, often at

a higher resolution. When fine-tuning, the patch size remains fixed, and the pre-trained position embeddings are resized via 2D interpolation.

The key advantage of ViT is that, by relying on self-attention rather than convolutions, they can capture global context more effectively. However, because they lack the strong, built-in inductive biases of CNN, ViT typically require large-scale pre-training to achieve competitive performance. Hybrid architectures that combine CNN with Transformer layers have also been proposed to integrate useful image-specific biases while leveraging the global modeling capabilities of self-attention.

### V.2.3 Transfer Learning and Neural Network Ensembles

Modern deep learning applications, especially in biomedical signal processing, face challenges such as limited data and the need for robust, reliable predictions. Researchers address these issues by applying transfer learning and ensemble methods. Transfer learning uses pre-trained models to overcome data scarcity, while ensemble methods combine multiple models to reduce variance and enhance generalization. Together, these strategies form a powerful framework for solving complex classification tasks under resource constraints.

#### Transfer Learning

Transfer learning (Caruana, 1997) leverages knowledge from a model trained on a large, general dataset (the base task) to improve performance on a different, often smaller target task. The central idea is that the features learned, especially in the lower layers of a CNN, capture generic patterns such as edges and textures that are useful across a variety of recognition tasks. By reusing these pre-trained features, transfer learning enables the training of target models with limited labeled data while mitigating overfitting (Yosinski et al., 2014).

Consider a pre-trained CNN defined as

$$f(\mathbf{x}; \boldsymbol{\theta}_c, \boldsymbol{\theta}_n), \tag{V.49}$$

where $\mathbf{x} \in \mathbb{R}^d$ represents an input image, $\boldsymbol{\theta}_c$ denotes the parameters of the convolutional base, capturing general visual features, and $\boldsymbol{\theta}_n$ represents the parameters of the classification head. Typically, the convolutional base is trained on a large-scale dataset such as ImageNet, and its learned representations are then transferred to a target task.

In a standard transfer learning approach, the convolutional base is kept fixed (frozen), preserving its learned representations, while the classification head is retrained on the

target data. This procedure can be formalized as

$$\boldsymbol{\theta}_n^* = \arg\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{N} \mathcal{L}\Big(\sigma\Big(\mathbf{W}_n\, f_c(\mathbf{x}_i; \boldsymbol{\theta}_c) + \mathbf{b}_n\Big),\, y_i\Big), \tag{V.50}$$

where $f_c(\mathbf{x}_i; \boldsymbol{\theta}_c)$ denotes the feature extraction performed by the pre-trained convolutional base, $\mathbf{W}_n$ and $\mathbf{b}_n$ are the weights and biases of the classification head, $\sigma(\cdot)$ is the activation function, $\mathcal{L}$ is the loss function, and $N$ is the number of target training samples.

When the target dataset is larger or the base and target tasks differ more substantially, it is often beneficial to fine-tune the model. In fine-tuning, a subset of the convolutional layers is unfrozen and updated along with the classification head, allowing the network to adjust its learned representations to the new task while still leveraging the robust, general features from the base task. This strategy can be expressed as

$$\{\boldsymbol{\theta}_c', \boldsymbol{\theta}_n'\} = \arg\min_{\boldsymbol{\theta}_c, \boldsymbol{\theta}_n} \sum_{i=1}^{N} \mathcal{L}\Big(f(\mathbf{x}_i; \boldsymbol{\theta}_c, \boldsymbol{\theta}_n),\, y_i\Big) \quad \text{subject to} \quad \|\boldsymbol{\theta}_c' - \boldsymbol{\theta}_c^{(0)}\| < \epsilon, \tag{V.51}$$

where $\boldsymbol{\theta}_c^{(0)}$ represents the original convolutional parameters, $\boldsymbol{\theta}_c'$ the fine-tuned parameters for the unfrozen layers, and the constraint ensures that the updated parameters remain close to the robust features initially learned.

The benefits of transfer learning are clear: by leveraging large-scale annotated datasets, it significantly reduces training time and improves performance on target tasks with limited data. However, when the target task differs substantially from the source task, the transferred features may not be entirely optimal, necessitating careful fine-tuning. Empirical studies have shown that transferring features from the lower layers, where representations are more generic, often results in superior performance (Donahue et al., 2014; Zeiler et al., 2014). Nonetheless, fine-tuning these lower layers involves adjusting a larger number of parameters, which increases the risk of overfitting if target dataset is small.

**Ensembles of Multiple Neural Networks**

Ensemble methods are strategies that enhance the performance and robustness of neural network models by combining the outputs of several individual models into a single prediction with improved accuracy and generalization. In CNN, such ensembles are particularly effective at mitigating issues like overfitting and sensitivity to initialization. A critical factor for successful ensembles is the promotion of diversity among the networks. This diversity can be achieved by training each CNN with different random initializations, on distinct subsets of data (bagging), or by varying hyperparameters and even network architectures.

A particularly intuitive method for forming ensembles is voting. For a given input $\mathbf{x}$,

suppose the $m$th CNN produces a probability estimate $p_m(y \mid \mathbf{x})$ for a target class $y$. The ensemble prediction is then computed as a weighted sum:

$$p_{\text{ens}}(y \mid \mathbf{x}) = \sum_{m=1}^{M} w_m \, p_m(y \mid \mathbf{x}), \tag{V.52}$$

where the weights $w_m$ satisfy $\sum_{m=1}^{M} w_m = 1$. In many cases, equal weighting ($w_m = 1/M$ for all $m$) is used.

In soft voting, each network outputs a probability distribution over the classes, and these distributions are averaged according to Equation (V.52). This approach smooths out uncorrelated errors among the models, resulting in reduced variance and more reliable, calibrated predictions. In contrast, hard voting requires each CNN to make a discrete class prediction, with the final decision determined by a majority vote. Although soft voting generally produces smoother probability estimates, hard voting can be beneficial when individual model outputs are noisy or exhibit poor calibration.

While ensembles of CNN offer significant improvements, they also introduce some practical challenges. Training multiple deep networks requires substantial computational resources and memory, and the inference process can incur additional latency, which may be critical in real-time applications. Snapshot ensembles alleviate this cost by training a single CNN and saving multiple snapshots at different convergence stages. These snapshots effectively form an ensemble with significantly reduced training overhead (Huang et al., 2017).

Beyond these standard methods, more advanced strategies such as expert gating and cascading provide dynamic approaches to ensembling. In expert gating, a dedicated network assigns input-dependent weights to specialized subnetworks (experts) within the ensemble, thereby emphasizing the contributions of the experts best suited for a given input. Cascading, on the other hand, organizes models into a sequential pipeline where a lightweight model first processes the input; if its output is uncertain or indicates a complex case, the input is then forwarded to a more sophisticated model. This sequential processing allocates computational resources efficiently by handling easy cases quickly and reserving detailed analysis for challenging instances. Both strategies can be further enhanced via meta-learning techniques, wherein a separate model learns to optimally combine the ensemble outputs.

The combination of transfer learning and neural network ensembles can create a synergistic framework that addresses both data scarcity and the need for robust, reliable predictions. This dual strategy is particularly effective when the source and target domains share similar low-level features but differ in high-level semantics. For example, in biomedical signal classification tasks, transfer-learned CNN have achieved accuracies as high as

92% on arrhythmia detection using relatively few samples per class (Isin et al., 2017). The ensemble approach improves uncertainty estimates and enhances the calibration of predictions, which is critical in safety-sensitive clinical environments (Lakshminarayanan et al., 2017). Together, transfer learning and ensembling can facilitate effective learning from limited data while enhancing the reliability of predictions, thus having a good potential for clinical applications.

### V.2.4 Data Augmentation Methods

Though deep neural networks, including CNN and Transformers, have achieved remarkable performance across various domains, their success depends heavily on the availability of large, diverse training datasets. In many practical applications, especially in biomedicine, the number of available samples is limited, making these models prone to overfitting. Overfitting occurs when a model learns to memorize the training data rather than capturing the underlying patterns, resulting in poor generalization to unseen examples. Data augmentation is a form of regularization designed to address this issue by artificially expanding the training set with realistic variations, thereby enabling the model to learn more robust and generalizable representations.

In supervised learning, we assume that data samples $(\mathbf{x}, y)$ are drawn from an unknown true distribution $P(\mathbf{x}, y)$. The goal is to find a function $f$ that minimizes the true risk

$$R(f) = \int \mathcal{L}\big(f(\mathbf{x}), y\big)\, dP(\mathbf{x}, y), \tag{V.53}$$

where $\mathcal{L}$ is a loss function measuring the discrepancy between the prediction $f(\mathbf{x})$ and the true target $y$. However, since $P(\mathbf{x}, y)$ is unknown, a common approach is to approximate it using the finite set of training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$.

A standard approximation is given by the empirical distribution,

$$\hat{P}(\mathbf{x}, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(\mathbf{x} - \mathbf{x}_i)\delta(y - y_i), \tag{V.54}$$

which places a point mass, that is a Dirac delta measure, at each training point, creating a discrete probability distribution. Accordingly, the Empirical Risk Minimization (ERM) principle (Vapnik et al., 1998) seeks to minimize the empirical risk,

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\big(f(\mathbf{x}_i), y_i\big), \tag{V.55}$$

where $f(\mathbf{x}_i)$ is the prediction for the $i$th sample and $y_i$ is the corresponding true target. This approach approximates the true risk, defined as an expectation over the unknown distribution $P(\mathbf{x}, y)$, by summing the losses over the training data points.

Classical learning theory guarantees the convergence of ERM under the condition that the capacity of the learning machine, often quantified by the VC-complexity (Vapnik et al., 2015), remains bounded as the number of training samples increases. In modern deep learning, however, models with a large number of parameters can easily exceed the number of training examples, which may result in overfitting and drastic changes in predictions for samples that are only slightly different from those seen during training (Zhang, 2017).

To mitigate these issues, the Vicinal Risk Minimization (VRM) principle (Chapelle et al., 2000) extends ERM by considering a *vicinity* around each training sample. Rather than approximating the true data distribution with point masses, VRM constructs a continuous *vicinal* distribution defined via a vicinity function $\nu(\tilde{\mathbf{x}}, \tilde{y} \mid \mathbf{x}_i, y_i)$. This function acts as a probability density over the neighborhood of each training sample $(\mathbf{x}_i, y_i)$ and may, for example, be instantiated as a Gaussian kernel (which yields Gaussian noise injection augmentation) (Zhang, 2017). The vicinal distribution is defined as

$$P_v(\tilde{\mathbf{x}}, \tilde{y}) = \frac{1}{n} \sum_{i=1}^{n} \nu(\tilde{\mathbf{x}}, \tilde{y} \mid \mathbf{x}_i, y_i). \tag{V.56}$$

By drawing virtual samples $(\tilde{\mathbf{x}}, \tilde{y})$ from this smoothed distribution, one effectively augments the training data. This augmentation yields smoother decision boundaries and encourages the model to behave more linearly in the regions between the observed examples.

A straightforward data augmentation strategy is to apply geometric transformations that preserve the semantic content of an image while modifying its spatial configuration. For example, affine transformations – such as rotations, scaling, and translations, can be represented as

$$\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{V.57}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input image, $\mathbf{A}$ is a transformation matrix (which may encode rotation by an angle $\theta$ and scaling by a factor $s$), and $\mathbf{b}$ is a translation vector. For instance, a rotation combined with scaling is expressed by

$$\mathbf{A} = \begin{bmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{bmatrix}. \tag{V.58}$$

In addition to spatial modifications, photometric transformations alter the appearance of an image by modifying its brightness, contrast, or by adding noise. A common technique is *Gaussian noise injection*, in which the input is perturbed by noise sampled from a

Gaussian distribution:

$$\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}). \tag{V.59}$$

Such noise addition helps the network learn smoother functions, thereby reducing sensitivity to small perturbations and preventing overfitting.

More advanced methods construct new training examples by combining or interpolating between existing ones. One notable approach is *MixUp* (Zhang, 2017), which generates virtual examples by forming convex combinations of pairs of samples. Given two training examples $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$, the mixed sample is defined as

$$\tilde{\mathbf{x}} = \lambda\, \mathbf{x}_i + (1 - \lambda)\, \mathbf{x}_j, \quad \tilde{y} = \lambda\, y_i + (1 - \lambda)\, y_j, \tag{V.60}$$

where the mixing coefficient $\lambda$ is drawn from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$, with a typical choice of $\alpha = 0.2$. This augmentation strategy is motivated by the idea that linear interpolations in the feature space should correspond to linear interpolations in the label space, thus enforcing a linear behavior between training examples. This linearity serves as an inductive bias that can suppress undesirable oscillations in the model's predictions outside the training set.

*Random Erasing* (Zhong et al., 2020) simulates *occlusion* – a phenomenon in computer vision where parts of an object are hidden or blocked from view. In data augmentation, occlusion is intentionally introduced by randomly selecting a rectangular region $\mathcal{R}$ within an image and replacing the pixel values in that region with either random noise or a constant value (often the mean pixel value). Here, $\tilde{\mathbf{x}}[u, v]$ denotes the pixel value at the coordinate $(u, v)$ in the augmented image, and $\mathcal{N}(\mu, \sigma^2)$ represents a sample drawn from a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. This process is described as

$$\tilde{\mathbf{x}}[u, v] = \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad (u, v) \in \mathcal{R}. \tag{V.61}$$

By forcing the network to rely on multiple and spatially distributed cues rather than specific local features, random erasing improves robustness to occlusions and enhances the model's ability to generalize.

Another method commonly used for sequential data is *Temporal Shifting*. This technique cyclically shifts a sequence to introduce invariance to time delays. If $\mathbf{x} \in \mathbb{R}^T$ is a time series of length $T$, a shifted sequence is given by

$$\tilde{\mathbf{x}}[t] = \mathbf{x}\Big[(t + \Delta t) \bmod T\Big], \tag{V.62}$$

where $\Delta t$ is uniformly sampled from $[0, T)$. This approach is particularly effective for periodic signals, ensuring that the model is robust to phase variations.

Another data augmentation strategy involves occluding a randomly selected rectangular region of the input, spanning both the X and Y dimensions of a spectrogram or image. In computer vision, this approach is known as the *Cutout* method, in which a contiguous patch of an image is removed to create a partially occluded sample (DeVries et al., 2017). A similar principle is applied in spectrogram augmentation through time-frequency masking (Park et al., 2019). By training on inputs with such missing regions, the model is encouraged to rely on broader contextual information rather than focusing solely on specific local features, thereby enhancing robustness to occlusions and noisy or missing data. Empirical studies have demonstrated the efficacy of this strategy; for example, masking out blocks of time steps or frequency channels in audio spectrograms has led to significant improvements in speech recognition performance (Park et al., 2019), and employing Cutout-style occlusions in images has been shown to act as an effective regularizer that reduces test error by promoting better generalization.

Data augmentation is thus a powerful regularization tool that expands the effective training distribution by generating virtual samples that are similar to the original examples. By incorporating prior knowledge about the data, augmentation methods mitigate the risks of overfitting – a critical consideration in fields where data scarcity is a persistent challenge.

## V.3 Skeletal Muscle Disorders Datasets: Classification

In this section we investigate our hypothesis that iEMG recordings contain class-specific patterns that can distinguish sodium- from chloride-channel myotonias. We proceed in two stages. First, we follow a feature-engineering route: the scalar time-domain descriptors extracted in Chapter IV are used for classification with Logistic Regression, SVM, Random Forests and XGBoost. A patient-grouped five-fold cross-validation ensures that no subject appears in both training and test partitions. To identify redundancy and assess feature importance, we perform correlation heatmap analyses.

In the second stage we set aside time-domain statistics and analyze the scalograms, obtained in Chapter IV directly. Treated as images, they are passed through ImageNet-pre-trained backbones. We benchmark a range of loss functions, data augmentation schemes and architectures, then ensemble the two top performers to quantify the gain over single networks and assess whether the resulting probabilities are reliable.

To interpret the deep models' decisions, we apply Gradient-weighted class activation mapping to the InceptionResNetV2 network trained on Morse-wavelet scalograms. Saliency maps are generated for both the measured and the simulated channelopathy datasets, as well as for an external fibrillation potentials corpus to identify the signal morphologies the network deems most discriminative and to discuss their physiological plausibility. Finally,

the fibrillation dataset is processed with the same pipeline as the channelopathy data to test our supplementary hypothesis that the method can be applicable across neuromuscular pathologies.

### V.3.1 Results of Classical Machine Learning Methods

We extracted patient IDs to define groups for cross-validation, ensuring that all samples from the same patient are exclusively assigned to either the training or test set. Using scikit-learn's `GroupKFold` with 5 splits, we trained four classifiers: Logistic Regression, SVM, Random Forest, and XGBoost. For each fold, evaluation metrics (accuracy, recall, precision, F1 score, specificity, ROC-AUC, average precision, Cohen's Kappa, and balanced accuracy) were computed and then averaged across folds. Additionally, confusion matrices were summed over all folds to produce an averaged confusion matrix for plotting.

In our implementation of the classification methods, standard machine learning models were employed with targeted modifications to enhance convergence and robustness. The Logistic Regression model, for example, was configured to allow up to 2000 iterations to ensure convergence. The SVM was integrated into a pipeline that applied feature scaling using a standard scaler and enforced balanced class weights, with probability estimates enabled. Within each cross-validation fold, hyperparameter tuning was performed using an inner 3-fold search to determine the optimal parameters, including kernel type. The Random Forest classifier was implemented with 100 trees and a fixed random state for reproducibility, while the XGBoost classifier was configured with settings that disable automatic label encoding and specify a particular loss metric, also employing a fixed random seed to ensure consistency.

We computed the correlation matrix on the full dataset, prior to forming splits, to capture the complete set of pairwise relationships among numeric features. The correlation matrix is a symmetric table where each entry represents the Pearson correlation coefficient between two features. Values near 1 indicate a strong positive linear relationship, values near -1 indicate a strong negative relationship, and values around 0 suggest little to no linear correlation. This analysis is useful for identifying redundant or noisy features that might adversely affect model performance.

To enhance interpretability, we visualized the correlation matrix as a heatmap. In this representation, color intensity reflects both the magnitude and direction of the correlations – warmer colors indicate strong positive correlations, cooler colors indicate strong negative correlations, and lighter shades represent weak correlations. The diagonal cells, which are always perfectly correlated (coefficient of 1), are excluded from interpretation. Our heatmap (Fig. V.18), computed using all available samples prior to any train/test split, reveals distinct clusters of highly correlated features. For example, RMS, MAV, and VAR form a prominent block of warm colors, suggesting that they capture similar signal

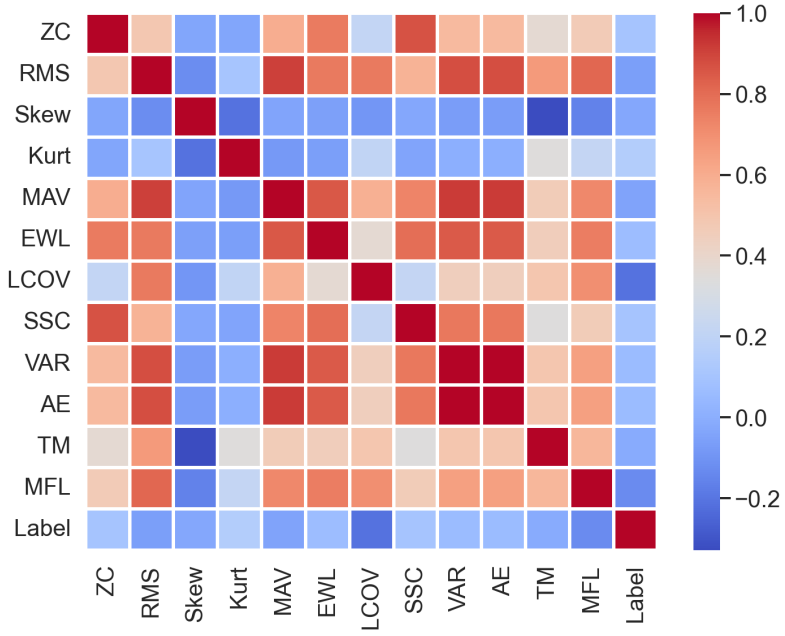



**Figure V.18:** Correlation heatmap computed on the entire merged EMG dataset, showing pairwise Pearson correlation coefficients among features.

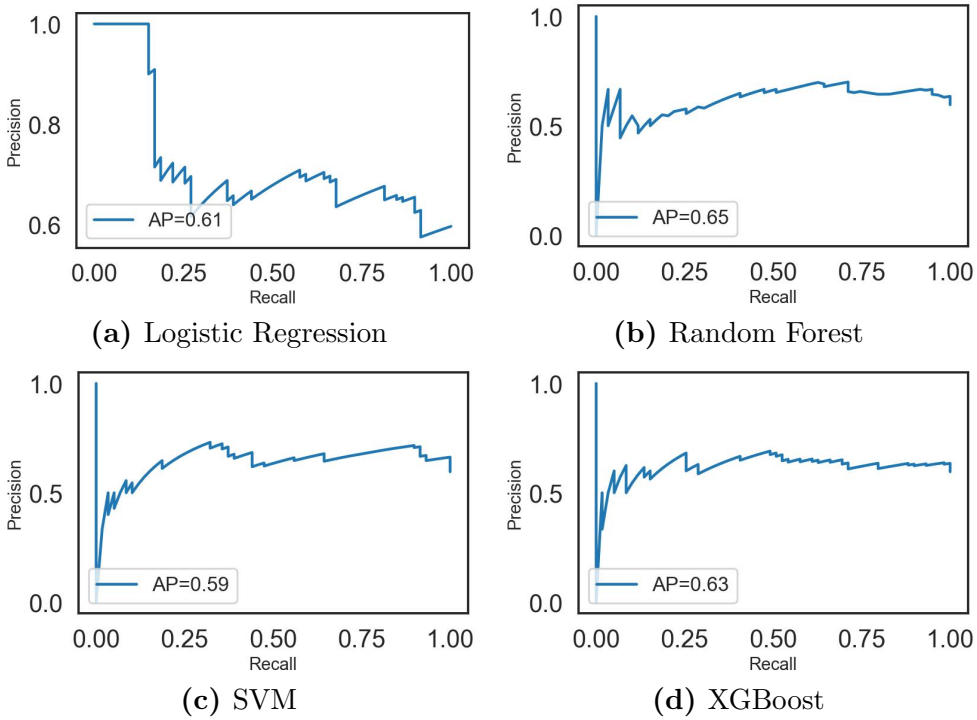
properties and may be redundant. In contrast, features such as skewness with TM and skewness with kurtosis display lower correlation coefficients, implying that these features capture distinct aspects of the data that are valuable for prediction.

Moreover, additional patterns emerge from the analysis. Certain feature pairs, such as ZC with SSC and VAR with average energy (AE), exhibit very high correlations, which further indicates redundancy. Notably, LCOV emerges as the strongest predictor of the labels, followed by MFL, while kurtosis appears to be the weakest. These insights not only deepen our understanding of the underlying data structure but also guide our feature selection process for subsequent analyses.

A precision-recall curve is a graphical tool used to assess classifier performance, particularly in imbalanced datasets. It illustrates the trade-off between precision and recall across a range of decision thresholds. For each sample, the classifier outputs a probability or score, and by varying the threshold from 0.0 to 1.0, different precision- recall pairs are obtained and plotted. When a high threshold is applied, the model only labels highly confident samples as positive, resulting in high precision but low recall; lowering the threshold increases recall at the expense of precision. Ideally, a model will produce a curve that stays near the top-right corner, signifying both high precision and high recall, whereas a model with no predictive skill would yield a nearly horizontal line at the overall positive rate.

In our experiments, we generated Precision–Recall curves for each classifier to compare

their performance. The results show that while none of the models drastically outperforms the others, the Random Forest model achieves a slightly higher average precision of approximately 0.65, with the other models clustering between 0.59 and 0.63. This pattern indicates that the dataset poses a challenging classification problem, as evidenced by the inherent trade-off between precision and recall.



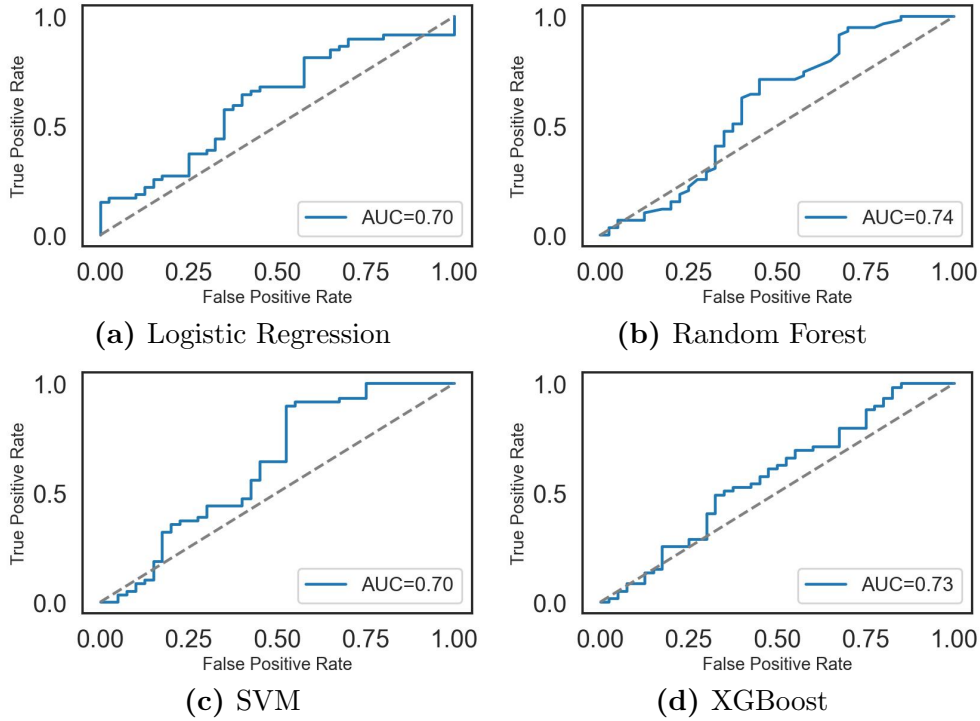

**(a)** Logistic Regression **(b)** Random Forest

**(c)** SVM **(d)** XGBoost

**Figure V.19:** Precision-recall curves of four different classical machine learning models.

Examining the ROC curves (Fig. V.20) for the different classifiers reveals that the Random Forests achieves the highest AUC at approximately 0.744, followed by the XGBoost at about 0.731, with both logistic regression and the support vector machine around 0.697. Although all models perform better than random guessing, the performance gap between the best and the weakest is modest, and none of the ROC curves approach the ideal top-left corner – a hallmark of a near-perfect classifier. In a clinical setting, where highly reliable detection of positive cases is critical, one would expect a ROC curve that remains close to the upper and left boundaries, indicating both high sensitivity and a low rate of false positives.

When these ROC results are considered alongside the precision-recall curves, it becomes evident that while the models can distinguish between classes to some extent, the overall balance between correctly identified positives and the cost of false positives or negatives remains suboptimal for clinical diagnostic purposes. Although an AUC between 0.70 and 0.74 might be acceptable in certain applications, clinical contexts generally demand higher performance to minimize the risk of overlooking critical cases.



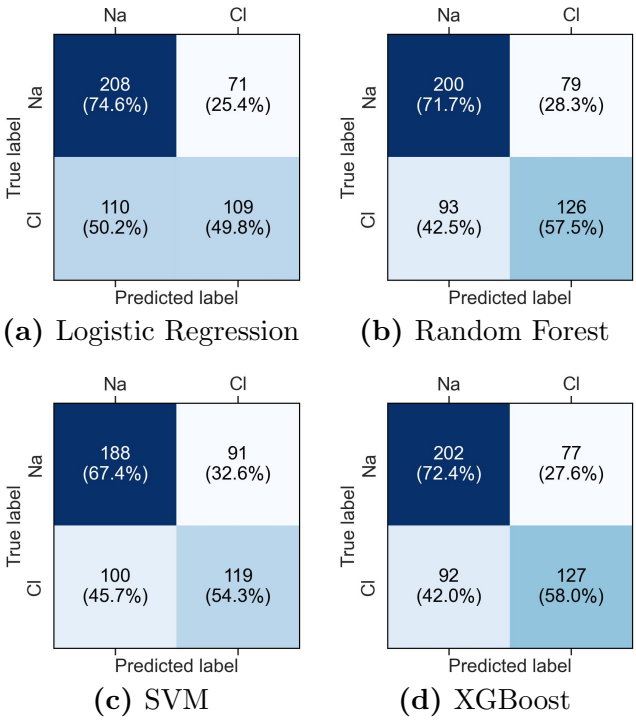

**(a)** Logistic Regression **(b)** Random Forest

**(c)** SVM **(d)** XGBoost

**Figure V.20:** ROC curves of four different classical machine learning models.

The confusion matrices, Fig. V.21, further illustrate the performance differences among the classifiers. Logistic regression tends to correctly identify negative cases but misses a larger proportion of positive instances, while the Random Forests and XGBoost capture more positives and exhibit fewer false negatives. The SVM falls between these extremes. These observations are consistent with the trends noted in the precision-recall and ROC analyses, where Random Forest and XGBoost showed a slight edge. In a clinical context, minimizing both false negatives and false positives is crucial because overlooking positive cases can have serious consequences. However, none of the models delivers a confusion matrix that is clearly superior for high-stakes diagnostic applications, as each still misclassifies a significant portion of the data.

When examining the grouped classification metrics for logistic regression, random forest, SVM, and XGBoost, Tab. V.1, it becomes evident that all four methods exhibit moderate, yet suboptimal, performance. Notably, the tree-based and boosting-based approaches demonstrate slightly higher overall accuracy, F1 scores, and other predictive measures, while logistic regression and SVM lag somewhat behind. These findings align with our earlier observations from the precision-recall curves, ROC curves, and confusion matrices, which indicated that random forest and XGBoost more effectively balance the trade-off between capturing true positives and limiting false positives and negatives. Nonetheless, none of the models achieves the high levels of sensitivity, specificity, and precision typically required for a reliable medical diagnostic test.




**(a)** Logistic Regression **(b)** Random Forest

**(c)** SVM **(d)** XGBoost

**Figure V.21:** Confusion matrices of four different classical machine learning models.

**Table V.1:** Metrics for Classical Machine Learning Models on Full Feature Set

| Classifier | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 64 | 50 | 74 | 60 | 53 | 0.70 | 61 | 0.23 | 62 |
| Random Forest | 65 | 60 | 72 | 62 | 58 | 0.74 | 65 | 0.30 | 66 |
| SVM | 62 | 53 | 67 | 56 | 49 | 0.70 | 59 | 0.18 | 60 |
| XGBoost | 66 | 61 | 73 | 62 | 59 | 0.73 | 63 | 0.30 | 67 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; AP = Average Precision; Bal. Acc. = Balanced Accuracy.

For our subsequent analysis, we excluded the redundant features VAR and ZC based on our correlation heatmap. Table V.2 summarizes the metrics obtained with the reduced feature set. In comparison to the full feature set, we do not observe a noticeable improvement in performance or stability. In the full feature scenario, linear classifiers can suffer from multicollinearity, leading to less robust coefficient estimates and diminished discriminative power. However, after excluding VAR and ZC, the performance of Logistic Regression remains approximately the same as before.

Similarly, tree-based models such as Random Forest and XGBoost maintained or even slightly worsened their average precision and overall accuracy with the reduced feature set. Although these models are inherently resilient to redundant inputs and can benefit from a

cleaner feature space that mitigates overfitting and improves interpretability, in our case, removing the redundant features did not enhance their performance. The only model that showed improvement was the SVM. Previously hindered by an overly conservative decision threshold, the SVM demonstrated improvements across all metrics, except for ROC AUC, following the removal of the redundant features.

In conclusion, reducing the feature set did not compromise our models' ability to capture the underlying data structure; indeed, for the SVM, the metrics confirm that VAR and ZC were redundant. The refined feature set yielded a more parsimonious SVM model with modest performance gains and enhanced stability. Nonetheless, the metrics for all models remain insufficient for the pipeline to be deployed in a clinical setting. Our experiments with classical methods based on hand-crafted features reveal that these approaches struggle to capture the full variability of EMG signals, leading to suboptimal classification performance.

**Table V.2:** Metrics for Classical Machine Learning Models on Reduced Feature Set

| **Classifier** | **Acc.** | **Rec.** | **Spec.** | **Prec.** | **F1** | **ROC AUC** | **AP** | $\kappa$ | **Bal. Acc.** |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 63 | 49 | 74 | 58 | 52 | 0.69 | 62 | 0.22 | 62 |
| Random Forest | 61 | 57 | 72 | 60 | 57 | 0.74 | 65 | 0.27 | 65 |
| SVM | 64 | 57 | 72 | 57 | 51 | 0.62 | 55 | 0.21 | 62 |
| XGBoost | 65 | 58 | 74 | 62 | 58 | 0.75 | 64 | 0.29 | 66 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; AP = Average Precision; Bal. Acc. = Balanced Accuracy.

While the models can distinguish between classes, they require further refinement to achieve close to clinical-grade reliability. For medical AI and Software as a Medical Device, regulators do not impose universal performance cut-offs. Instead, acceptance criteria for example, sensitivity and specificity must be tailored to the intended clinical use, pre-specified in the study protocol, and demonstrated on independent test sets with two-sided 95% confidence intervals (International Medical Device Regulators Forum (IMDRF), 2017; U.S. Food and Drug Administration, 2007). In regulatory practice, performance targets are typically defined so that the *lower* bound of the 95% confidence interval meets or exceeds the acceptance threshold; for example, the FDA-cleared autonomous IDx-DR system for diabetic retinopathy prospectively targeted $\geq$85.0% sensitivity and $\geq$82.5% specificity, successfully achieving these goals (U.S. Food and Drug Administration, 2018).

Our evaluation of classical machine learning methods shows that, while they exhibit discriminative power exceeding that of random classifiers, their limitations highlight the need to explore more advanced approaches. Deep neural networks, in particular, offer

the potential to extract richer and more informative features, thereby delivering the performance improvements necessary for clinical deployment.

### V.3.2 General Pipeline: Transfer Learning and Ensembles

In the previous subsection, we demonstrated that classical machine learning models trained on time-domain features extracted from the skeletal muscle channelopathy dataset were able to distinguish between different types of channelopathies, though only to a limited extent. However, the performance of these models was insufficient for clinical application. Nonetheless, these initial results provided evidence supporting our hypothesis (H1), indicating that skeletal muscle signals do indeed contain discriminative features.

In Chapter IV, we processed these signals using CWT, enabling the extraction of richer and more descriptive features. Since CWT outputs are represented as image-like structures known as scalograms, a natural approach was to classify these images using deep neural networks initially developed for ImageNet classification tasks. Such networks have the capability to efficiently extract meaningful and discriminative features from image-based data.

However, ImageNet-trained networks typically require large datasets for effective training, which presents a challenge given the limited size of our skeletal muscle channelopathy dataset (approximately 500 samples). To address this limitation, our strategy primarily involves transfer learning and model ensembling techniques. Both approaches aim to maximize classification performance while mitigating challenges associated with data scarcity, class imbalance, and potential overfitting.

Since our ultimate objective is to develop an automated pipeline (hypothesis H2 for the channelopathy dataset) suitable for clinical applications (hypothesis H3 for the channelopathy dataset), we prioritize practical, robust, and computationally efficient methods. These methods should require minimal data processing and parameter tuning and be adaptable to other measurement types (extending hypothesis H2 to the fibrillation potentials dataset). To identify such optimal methods, we conduct extensive benchmarking tests, evaluating various ImageNet-based architectures, augmentation strategies, and analysis methods through grouped cross-validation. The performance criterion for selecting the best models is based primarily on achieving the lowest validation loss, with further metrics computed for validation and test sets during inference.

In this subsection, we provide a comprehensive description of our classification methodology and present experimental results obtained using this approach. Code for signal processing, model training, and evaluation will be made available at `https://github.com/morphoemilie/Classification_of_iEMG`.

### Description of Single Model Training

In the transfer learning framework, the weights of a pre-trained neural network are kept frozen during training to leverage pre-learned representations effectively. We chose InceptionResNetV2 as our baseline model due to its exceptional performance in preliminary experiments conducted with synthetic data (these results will be summarized in Subsection V.3.4).

Specifically, the pre-trained InceptionResNetV2 model serves as a fixed feature extractor. Its original classification head is removed, leaving only convolutional layers and a global average pooling layer, which generates a 1536-dimensional feature vector. Subsequently, this feature vector undergoes further processing within a custom classification head. Initially, dropout regularization is applied to prevent overfitting. Then, the vector passes through a fully connected layer reducing its dimensionality to 1024, activated by a ReLU function. Another fully connected layer compresses these representations further to 512 features, also activated by ReLU. Finally, an output layer maps the 512-dimensional representation to the desired class predictions, completing the robust and efficient classification pipeline.

To ensure consistent comparability across various models and methods, we utilize evaluation metrics identical to those used in the classical machine learning experiments, supplemented with calculations of model entropy and confidence. Although additional metrics such as accuracy and ROC-AUC are computed for comprehensive analysis, the validation loss is continuously monitored during training to determine optimal model checkpoints within each cross-validation split. Extensive experimentation identified that the best performance is consistently achieved using the AdamW optimizer coupled with a cosine annealing scheduler, alongside an early stopping mechanism with a patience parameter set to 7 epochs.

Initially, our deep neural network pipeline involves assessing individual pre-trained ImageNet models as feature extractors. We systematically test various architectures to evaluate their respective performance, subsequently selecting the top-performing models for pipeline enhancement. Our initial experiments use minimal pre-processing, limited to standardization on scalograms generated using Morse wavelets with parameters gamma = 3 and P = 60, 12 voices per octave, and frequency coverage from 0 Hz up to 2000 Hz.

### Loss Function

To address class imbalance and enhance model generalization in our transfer learning pipeline, we bench-marked three distinct loss functions: Focal Loss, Weighted Cross-Entropy, and Label Smoothing. The averaged metrics over five splits, along with their standard deviations, are summarized in Tab. V.3.

Focal Loss (Lin et al., 2017) modifies the standard cross-entropy by introducing a scaling factor $(1 - p_t)^\gamma$, where $p_t$ is the predicted probability for the true class. This factor reduces the contribution of easily classified examples, placing greater emphasis on harder-to-classify cases, which makes Focal Loss particularly beneficial for datasets with extreme class imbalance. However, despite achieving the lowest validation loss in our experiments, the model trained with Focal Loss exhibited notably poorer generalization on the independent test set. Specifically, we observed decreased accuracy, recall, and F1-score, likely due to its excessive focus on challenging or noisy examples, resulting in overfitting to specific validation cases.

Weighted Cross-Entropy tackles class imbalance by assigning higher weights to under-represented classes proportional to their inverse frequency, thereby balancing their influence during training. Given our dataset's mild imbalance (approximately 1.26:1), this method successfully improved performance metrics on the test set compared to Focal Loss. Nevertheless, it did not achieve optimal results since it primarily addresses imbalance without providing additional regularization against overconfidence. Consequently, the generalization improvements were limited, as the model could still become overly confident about its predictions.

Label Smoothing (Szegedy et al., 2016) acts as a regularization technique by softening the target probability distributions. Instead of using strict one-hot encoding, label smoothing assigns slightly non-zero probabilities across all classes, preventing the model from becoming overly certain about any single prediction. Our experiments revealed that Label Smoothing provided the highest overall performance on the test set across all evaluated metrics. Its effectiveness lies in its implicit regularization capability, reducing overfitting and maintaining stable gradients even for confidently classified examples. Interestingly, although Label Smoothing does not explicitly handle class imbalance, it indirectly mitigated our dataset's slight imbalance through its regularizing effect.

In comparing these three methods, the experimental outcomes clearly highlighted the strengths and weaknesses of each approach. Focal Loss, while effectively targeting hard examples in heavily imbalanced contexts, became counterproductive due to excessive emphasis on difficult samples in our mildly imbalanced scenario. Weighted Cross-Entropy improved results by balancing class contributions but lacked mechanisms to prevent prediction overconfidence, limiting its generalization capabilities. Ultimately, Label Smoothing emerged as the most effective method, consistently maintaining balanced gradients, preventing premature convergence, and significantly enhancing generalization. This underscores the importance of choosing a loss function that not only addresses specific challenges like class imbalance but also supports robust model generalization.

**Table V.3:** Classification Metrics for Loss Functions on Test Set

| Loss Function | Acc. | Rec. | F1 | Prec. | $\kappa$ | Bal. Acc. | Conf. | Ent. |
|---|---|---|---|---|---|---|---|---|
| Focal Loss | 68.0 ± 0.06 | 68.0 ± 0.07 | 67.0 ± 0.06 | 69.0 ± 0.08 | 0.39 ± 0.11 | 69.0 ± 0.06 | 0.72 ± 0.06 | 0.54 ± 0.07 |
| Label Smoothing | 73.0 ± 0.06 | 73.0 ± 0.07 | 73.0 ± 0.07 | 73.0 ± 0.07 | 0.46 ± 0.15 | 73.0 ± 0.07 | 0.88 ± 0.05 | 0.30 ± 0.08 |
| Weighted Cross Entropy | 71.0 ± 0.07 | 71.0 ± 0.06 | 70.0 ± 0.08 | 73.0 ± 0.06 | 0.40 ± 0.12 | 70.0 ± 0.06 | 0.85 ± 0.06 | 0.32 ± 0.11 |

**Note:** Acc. = Accuracy; Rec. = Recall; F1 = F1 Score; Prec. = Precision; $\kappa$ = Cohen's Kappa; Bal. Acc. = Balanced Accuracy; Conf. = Confidence; Ent. = Entropy; All metrics are reported as mean ± standard deviation across the five test splits.

### Deep Neural Network Architecture Tests

We expanded our experiments to evaluate a range of deep neural network architectures. These architectures included EfficientNet, EfficientNetV2, Xception, ResNeXt-50, InceptionNext tiny, MaxVit small variants (with input resolutions of 384×384 and 224×224 pixels), MaxVit tiny, ConvNext Base, Swin Transformer (Swin2cr small), and NasNetLarge. Upon benchmarking these architectures, we observed that InceptionResNetV2 and both MaxVit small variants consistently delivered superior results. ResNeXt-50, Xception, and MaxVit tiny closely followed, indicating these models also effectively adapted their ImageNet-learned features to the scalogram domain. Both EfficientNet models and InceptionNext tiny exhibited moderate underperformance, while ConvNext Base, NasNetLarge, and Swin2cr small showed the poorest results.

By analyzing the relationship between model performance and network size (parameter count), we noticed that smaller models such as EfficientNet ($\sim$ 5–6 million parameters), InceptionNext tiny ($\sim$ 6–7 million), and EfficientNetV2 ($\sim$ 20–25 million) generally underperformed, with accuracy lower than 70 %, compared to medium-sized networks with around 25 million parameters, such as ResNeXt-50, Xception, and MaxVit tiny, which showed accuracy slightly above 70 %. The best-performing architectures – MaxVit small and InceptionResNetV2 have approximately 55 million parameters, suggesting that models smaller than this may lack sufficient capacity to capture important features from the EMG scalograms. Conversely, the largest networks, ConvNext Base and NasNetLarge, each with roughly 90 million parameters, showed performance comparable to a random classifier, likely indicating severe overfitting due to their excessive capacity relative to the limited dataset size. Their extensive complexity, while advantageous for large-scale natural image tasks, possibly can be detrimental when the pretrained features are poorly aligned with patterns in the scalograms of EMG signals.

In general, our dataset significantly differs from the natural images present in the ImageNet database used for pretraining these architectures. Our scalograms, derived

from biomedical measurements, contain time-frequency patterns rather than typical visual textures and shapes. This domain difference can pose challenges for effective feature transfer and domain adaptation. The top-performing architectures, specifically InceptionResNetV2 and MaxVit small, incorporate multi-scale feature extraction and attention mechanisms that might adapt more efficiently to capture these patterns, compared to other networks we tested.

Conversely, the poor performance of Swin2cr small can likely be attributed to its window-based self-attention design, which partitions input images into fixed-sized windows. This design might hinder the model's ability to capture continuous global patterns, which are crucial for analyzing EMG scalograms. Similarly, the relatively weak results of NasNetLarge and ConvNext Base might stem from their substantial model capacity and architectural biases toward natural image features. ConvNext Base, in particular, employs convolutional strategies that rely heavily on large receptive fields and spatial normalization techniques, which seem less optimal for our data. Consequently, their complexity became a disadvantage rather than an asset, exacerbating overfitting and limiting effective domain transfer.

Based on this analysis, we selected InceptionResNetV2 and MaxVit small as the most suitable architectures for further integration into an ensemble model.

#### Application of Data Augmentation Methods

The baseline model, trained without any augmentation, established a reference with an accuracy of approximately 80 % on the validation and 71.32 % on the test set along with well-balanced precision, recall, and F1 scores. This baseline performance provided a reference point for evaluating the effectiveness of different data augmentation methods. While the model performed adequately under standard training conditions, there remained potential for improvement through data augmentation strategies. The averaged metrics obtained by testing various augmentation methods are summarized in Tab. V.4.

Injecting Gaussian noise resulted in a significant reduction in performance, with accuracy dropping to approximately 50 %. This decline indicates that the introduced noise affected critical signal features, thereby impeding the model's ability to extract meaningful patterns. In contrast, geometric transformations largely preserved the underlying data structure and achieved results comparable to the baseline on the validation set. However, these transformations negatively impacted test set performance, suggesting that they introduce insufficient variability to improve the model's generalization capabilities.

The individual application of MixUp and Random Erasing yielded modest yet noticeable improvements. Each method increased the validation set accuracy to approximately 81.5 % and enhanced metrics such as ROC AUC and average precision. However, when

performance was evaluated on the test set, Random Erasing demonstrated a more substantial improvement by achieving nearly 74 % accuracy and the highest Cohen's Kappa (approximately 0.47) among all models, whereas MixUp exhibited diminished performance relative to the baseline. Moreover, Random Erasing stands out with relatively balanced recall, precision, and F1 scores of approximately 74 %, a strong ROC AUC of about 0.82. Its robust Cohen's Kappa value indicates stronger agreement between predictions and ground truth than the baseline.

A more complex picture emerges when examining the combinations of augmentation method applied together. The simultaneous use of MixUp and Random Erasing did not enhance performance and, in some instances, slightly diminished overall results, suggesting that these two augmentations likely introduce redundant perturbations. Conversely, combining Cutout with Random Erasing yielded improved results, increasing accuracy to approximately 73 % on the test set. This finding might imply a synergistic relationship between Cutout and Random Erasing, although the performance gains were very modest. The most extensive composite of Cutout, MixUp, and Random Erasing combined, delivered a competitive test accuracy of 73.65 % and recorded the highest test-set $\kappa$ (0.509) and ROC-AUC (0.855), but these gains were modest relative to model complexity.

Among all evaluated augmentation techniques, time-shifting resulted in the highest performance metrics on the validation set. However, when tested on unseen data, its accuracy was lower than that achieved by Random Erasing, although it still surpassed the baseline model performance. This finding suggests that augmenting the dataset, even by shifted samples, generally enhances the predictive capability of the model. The high confidence and low entropy observed with time-shifted predictions further indicate that the model becomes more decisive and reliable when trained with this augmentation. Combining time-shifting with MixUp diminished these improvements on the validation set, while did not affect the results on the test set. This discrepancy could indicate that the validation set is more susceptible to overlapping perturbations, resulting in localized overfitting effects that are mitigated by the greater diversity in the test set.

These findings highlight the importance of selecting data augmentation methods that improve model generalization. Both Random Erasing and its combination with Cutout appear to introduce perturbations that enhance the model's learning capability without distorting the underlying signal, thereby facilitating robust classification. In contrast, augmentations such as Gaussian noise, and certain combinations like MixUp with Random Erasing, can disrupt critical features and degrade performance. The differences observed between validation and test outcomes highlight the risk that some augmentation methods may lead to overfitting, ultimately reducing overall model effectiveness.

Considering the intrinsic variability of iEMG measurements, particularly in the context

of myotonic discharges in patients with channelopathies, the superior performance of Random Erasing may stem from its ability to realistically simulate signal occlusions or missing segments. This augmentation strategy likely encourages the model to focus on robust and generalizable feature representations, improving its resilience to local variability. In contrast, overly disruptive methods might compromise critical diagnostic cues, leading to reduced predictive accuracy and miscalibrated model confidence.

The consistent performance of Random Erasing across key test set metrics, along with its ability to enhance model generalization, makes it a compelling choice for classifying iEMG signals. Consequently, Random Erasing will serve as the primary augmentation method in our model.

**Table V.4:** Validation and Test Metrics for Data Augmentation Methods in iEMG Signal Classification Averaged over Five Splits

| **Method** | **Set** | **Acc.** | **Recall** | **F1** | **Prec.** | $\kappa$ | **Bal. Acc.** | **Conf.** | **Ent.** | **ROC AUC** |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | Val. | 79.97 | 79.97 | 79.51 | 79.87 | 0.59 | 79.97 | 0.80 | 0.46 | 0.88 |
| | Test | 71.32 | 71.25 | 70.76 | 71.75 | 0.42 | 71.18 | 0.88 | 0.29 | 0.78 |
| Gaussian Noise | Val. | 55.23 | 100.00 | 71.16 | 55.23 | 0.00 | 50.00 | 1.00 | 0.00 | 0.50 |
| | Test | 58.73 | 52.86 | 49.44 | 55.19 | 0.02 | 50.94 | 1.00 | 0.00 | 0.51 |
| Geometric | Val. | 80.05 | 80.05 | 79.49 | 80.76 | 0.59 | 79.69 | 0.76 | 0.51 | 0.87 |
| | Test | 68.25 | 64.97 | 65.08 | 67.78 | 0.31 | 64.97 | 0.89 | 0.30 | 0.80 |
| Random Erasing | Val. | 81.50 | 81.50 | 80.85 | 82.19 | 0.62 | 81.20 | 0.76 | 0.51 | 0.89 |
| | Test | 74.05 | 74.62 | 73.87 | 74.79 | 0.47 | 74.23 | 0.75 | 0.53 | 0.82 |
| MixUp | Val. | 81.55 | 81.55 | 81.21 | 81.53 | 0.61 | 80.71 | 0.76 | 0.52 | 0.90 |
| | Test | 68.25 | 65.54 | 65.76 | 67.38 | 0.36 | 67.46 | 0.90 | 0.27 | 0.76 |
| Cutout | Val. | 80.85 | 80.85 | 80.46 | 81.80 | 0.61 | 80.52 | 0.80 | 0.47 | 0.89 |
| | Test | 66.67 | 68.19 | 66.63 | 67.86 | 0.43 | 72.25 | 0.87 | 0.32 | 0.81 |
| Time-shifting | Val. | 85.16 | 85.16 | 85.01 | 85.21 | 0.70 | 85.04 | 0.84 | 0.42 | 0.90 |
| | Test | 72.20 | 72.34 | 71.77 | 72.65 | 0.46 | 73.06 | 0.84 | 0.36 | 0.81 |
| Time-shifting MixUp | Val. | 80.19 | 80.19 | 79.32 | 80.39 | 0.59 | 80.24 | 0.75 | 0.53 | 0.90 |
| | Test | 72.30 | 72.54 | 72.04 | 72.69 | 0.47 | 73.60 | 0.85 | 0.37 | 0.81 |
| MixUp Rand. Eras. | Val. | 75.07 | 75.07 | 74.65 | 74.77 | 0.49 | 75.03 | 0.69 | 0.58 | 0.81 |
| | Test | 71.29 | 71.97 | 70.89 | 72.96 | 0.44 | 72.54 | 0.76 | 0.47 | 0.80 |
| Cutout + Rand. Eras. | Val. | 80.30 | 80.30 | 79.34 | 80.15 | 0.60 | 80.63 | 0.76 | 0.51 | 0.89 |
| | Test | 73.02 | 74.11 | 72.96 | 74.15 | 0.47 | 74.07 | 0.84 | 0.37 | 0.81 |
| Cutout + MixUp + Rand. Eras. | Val. | 79.56 | 79.56 | 79.17 | 79.29 | 0.59 | 79.56 | 0.76 | 0.52 | 0.89 |
| | Test | 73.65 | 74.38 | 73.45 | 74.81 | 0.51 | 75.94 | 0.84 | 0.36 | 0.86 |

**Note:** Acc. = Accuracy; F1 = F1 Score; Prec. = Precision; Conf. = Confidence; Ent. = Entropy; Bal. Acc. = Balanced Accuracy; Val. = Validation; Rand. Eras. = Random Erasing.

**Data Processing Tests**

In our experiments we systematically evaluated how different preprocessing techniques and parameter choices within wavelet transforms affect the classification performance. In the baseline configuration, we applied the Morse wavelet with parameters $\gamma = 3$ and $P = 60$ over a frequency range of 0–2000 Hz. This setting yielded an average accuracy of approximately 71.32 % with balanced class performance.

Filtering the signal prior to transformation improved the validation performance, which indicates that removing background noise and irrelevant components helps to clarify the discriminative features. However, the corresponding improvement in validation metrics did not carry over to the test set. In addition, targeted artifact removal, specifically the elimination of needle artifacts, resulted in a slight improvement in the test metrics over the baseline. Yet, analysis of the confusion matrices revealed that this improvement came at the cost of decreased sensitivity in detecting chloride channel defects. Such imbalanced classification performance raises concerns for clinical scenarios.

Altering the analyzed frequency ranges produced mixed outcomes. When the frequency band was restricted to exclude very low frequencies (0–20 Hz), the validation metrics were significantly enhanced, suggesting effective noise reduction. On the other hand, test performance deteriorated due to a reduced ability to detect the chloride class. This decline implies that the low-frequency band may contain essential discriminative information that is critical for accurate classification. A further narrowing of the frequency band to 20–1000 Hz generated intermediate results, yet it failed to produce meaningful improvements on the test set and resulted in increased imbalance in the classification metrics. Moreover, varying the wavelet parameter, which determines the resolution in time-frequency analysis, critically affected performance outcomes. Reducing P from 60 to 20 led to deteriorated validation and test set performances. This implies the importance of maintaining high temporal-frequency resolution for capturing informative features required for robust classification.

We also examined the impact of using the Morse wavelet at 24 voices per octave. Under this configuration, the model achieved about 73 % accuracy on the validation set with balanced precision and recall; however, the test set accuracy dropped to roughly 65 %. The corresponding confusion matrix showed that nearly half of the chloride examples were misclassified, thereby emphasizing the persistent challenge of attaining balanced class performance when tuning wavelet parameters.

Further experiments using the Morlet wavelet for transformation reinforced these findings. When applied over the frequency range of 0–2000 Hz, the Morlet wavelet produced test performance comparable to, but slightly lower than, that of the Morse wavelet. Although the exclusion of frequencies below 20 Hz modestly improved the validation accuracy, it

notably diminished the performance on the test set. This pattern suggests that while the elimination of low-frequency content may superficially enhance validation metrics, it also risks discarding valuable, generalizable information essential for robust testing.

Our experiments demonstrate that both Morse and Morlet wavelet analyses are sensitive to preprocessing and parameter choices. Although increasing the transformation resolution or altering frequency bands often improves performance metrics on validation sets, these improvements seldom generalize to unseen test data. Specifically, aggressive preprocessing techniques, such as filtering low frequencies or removing artifacts can inadvertently eliminate discriminative information or introduce bias, leading to overfitting and imbalanced class performance. Therefore, maintaining a broad frequency range and high time-frequency resolution is crucial for capturing generalizable features.

### A Single Model Summary

Based on evaluation of various neural network architectures, data augmentation strategies, and preprocessing techniques presented in the preceding sections, we developed an optimized single-model configuration. The experimental results demonstrated that the InceptionResNetV2 and MaxVit small architectures consistently delivered superior performance compared to other evaluated models. Consequently, we selected these two architectures, with input image sizes of 299 pixels for InceptionResNetV2 and 224 pixels for MaxVit small. Random Erasing was chosen as the primary data augmentation method, as it shows the best results on the test set.

The training process employed a transfer learning strategy, wherein the pre-trained neural networks served as fixed feature extractors. Model parameters were then optimized using an AdamW optimizer coupled with a cosine annealing learning rate scheduler and an early stopping mechanism triggered by validation loss. This training approach was systematically applied across multiple data splits to ensure both reliability and generalizability. A detailed overview of this pipeline is summarized in Algorithm 1.

### Ensemble of Best-performing Models

While the single-model configuration provides a strong baseline, integrating multiple model instances into an ensemble can further enhance robustness and improve generalization on unseen data. To exploit the complementary strengths of our best-performing architectures, we implemented an ensemble strategy. For each dataset split, four neural networks were independently trained: two instances of InceptionResNetV2 and two instances of MaxVit small. The selection of models for inclusion in the ensemble was based on their performance on the test set, ensuring the chosen instances exhibit high generalizability.

**Algorithm 1** Scalogram Image Classification Pipeline

**Require:** Dataset splits $\mathcal{D}$, scalogram images, and model specifications.
Load split directories and initialize logging, device, and timestamp.
**for** each split $s \in \mathcal{D}$ **do**
  Update configuration for split $s$.
  Load data: $(\mathcal{D}_{train}, \mathcal{D}_{val}) \leftarrow$ GET_DATA_LOADERS {Normalization and Random Erasing applied during loading.}
  Compute class weights from $\mathcal{D}_{train}$.
  Define loss function $\mathcal{L}$.
  Initialize model $\mathcal{M} \leftarrow$ GET_MODEL(`number_of_classes`, `dropout_rate`).
  Initialize optimizer and learning rate scheduler.
  **for** each epoch $e$ **do**
    **Train:** For each batch $(X, Y)$ in $\mathcal{D}_{train}$ {$X$ contains input scalogram image tensors and $Y$ the corresponding ground-truth labels.}
      Compute outputs using $\mathcal{M}$ and update parameters by minimizing $\mathcal{L}\big(\mathcal{M}(X), Y\big)$.
    **Validate:** For each batch $(X, Y)$ in $\mathcal{D}_{val}$ {$X$ and $Y$ are defined as above.}
      Compute outputs and loss, and update evaluation metrics.
    Save checkpoint if validation loss improves.
  **end for**
  Record and plot metrics for split $s$.
**end for**
Compute average metrics and aggregate confusion matrix across all splits.
Save final model, metrics, and performance plots.

Once the best-performing individual models were identified, they were integrated into an ensemble framework to classify scalogram images from independent test datasets. Ensemble predictions are generated by computing the softmax scores for every test sample from each model and subsequently averaging these scores to form a combined prediction for each class. The final class label for each test image is determined by selecting the class with the highest average score. To assess the effectiveness of this ensemble approach, multiple performance metrics are computed for each split, with the overall performance summarized by aggregating these metrics across splits via their mean and standard deviation. This evaluation demonstrates both the improved generalizability and the consistency of the ensemble across various subsets of data.

Taking advantage of the ensemble used for inference, we also quantify *predictive uncertainty* with the deep-ensemble approach of Lakshminarayanan et al. (2017). In this method, each independently-optimized network is viewed as a different draw from the space of plausible models learned from the data; averaging their soft-probabilities therefore approximates the Bayesian posterior predictive distribution, while the spread among members is a practical proxy for *epistemic* (model) uncertainty.

For our scalogram-based classifier the intuition is straightforward: when all four constituent

models assign most of their probability mass to the same class, the ensemble posterior is sharp and of low entropy, indicating that the input lies well inside the training manifold and is consequently likely to be labeled correctly. If the models disagree, the ensemble distribution becomes flatter, entropy rises and confidence falls, reliably flagging borderline, noisy, or out-of-distribution activations.

We quantify these effects with three proper scoring rules: (i) the *negative log-likelihood* (NLL) is the average −log probability assigned to the true class; it penalizes over-confident errors, attaining its minimum when predicted probabilities match empirical frequencies; (ii) the *Brier score* equals the mean-squared distance between the one-hot target vector and the predicted probability vector; it separates into a calibration component, which gauges how closely predicted probabilities match observed frequencies, and a refinement component, which captures how concentrated the distribution is, so lower scores signal predictions that are simultaneously well-calibrated and sharp (Brier, 1950); (iii) The *expected calibration error* (ECE) bins test samples by predicted confidence and reports the absolute gap between accuracy and confidence in each bin (Guo et al., 2017); a perfectly calibrated classifier has ECE = 0.

Let $M$ be the number of networks in the ensemble and let $p^{(m)}(y \mid x)$ denote the class posterior assigned by model $m$ to an input $x$. Averaging these posteriors gives the ensemble predictive distribution

$$\bar{p}(y \mid x) = \frac{1}{M} \sum_{m=1}^{M} p^{(m)}(y \mid x). \tag{V.63}$$

which can be viewed as a Monte-Carlo approximation of the Bayesian posterior predictive implicit in the stochastic training procedure. From $\bar{p}$ we derive two scalar uncertainty indicators – the *confidence* $c(x)$ and *predictive entropy* $H(x)$:

$$c(x) = \max_{y} \bar{p}(y \mid x), \qquad H(x) = -\sum_{y} \bar{p}(y \mid x) \log \bar{p}(y \mid x). \tag{V.64}$$

To assess how well these probabilities match reality we compute three proper scoring rules on a test set $\{(x_i, y_i)\}_{i=1}^{N}$:

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log \bar{p}(y_i \mid x_i), \tag{V.65}$$

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \big(1_{\{y_i = y\}} - \bar{p}(y \mid x_i)\big)^2, \tag{V.66}$$

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \big|\mathrm{acc}(b) - \mathrm{conf}(b)\big|. \tag{V.67}$$

where $n_b$ is the number of predictions in confidence bin $b$, $\mathrm{acc}(b)$ is their empirical accuracy, and $\mathrm{conf}(b)$ their mean confidence. Low values of NLL, Brier score, and ECE signal that ensemble probabilities are both sharp and well calibrated. Meanwhile, the empirical distributions of $c(x)$ and $H(x)$ reveal where the classifier is genuinely certain about its predictions and where high epistemic disagreement advises caution.

The next subsection will present a detailed analysis of the experimental results obtained using the pipelines described above. We will introduce our final results obtained over all of our datasets, which include simulated and measured data for skeletal muscle channelopathy, and also the results obtained by applying our final pipeline to the alternative problem of fibrillation potentials. It will discuss both single-model and ensemble approaches in terms of their classification performance, robustness, and potential for clinical application in EMG signal analysis.

**Algorithm 2** Ensemble inference and uncertainty evaluation

**Require:** Dataset splits $\mathcal{S}$, ensemble checkpoints $\{\mathcal{M}_k\}_{k=1}^{4}$, test loaders $\{\mathcal{L}_k\}$.
**for** each split $s \in \mathcal{S}$ **do**
  **Load** the four best checkpoints for split $s$ and set models to evaluation mode.
  Initialise lists for per-model logits $\mathcal{Z}_k$, true labels $\mathcal{Y}$ and file names $\mathcal{F}$.
  **for** each model $\mathcal{M}_k$ with loader $\mathcal{L}_k$ **do**
    **for** batch $(\mathbf{X}, \mathbf{y}, \mathbf{f})$ in $\mathcal{L}_k$ **do**
      $\mathbf{z} \leftarrow \mathcal{M}_k(\mathbf{X})$ {logits}
      Append $\mathbf{z}$ to $\mathcal{Z}_k$; append $\mathbf{y}$ to $\mathcal{Y}$; append $\mathbf{f}$ to $\mathcal{F}$.
    **end for**
  **end for**
  $\bar{\mathbf{z}} \leftarrow \frac{1}{4}\sum_{k=1}^{4} \mathrm{concat}(\mathcal{Z}_k)$ {logit averaging}
  $\hat{\mathbf{p}} \leftarrow \mathrm{softmax}(\bar{\mathbf{z}})$; $\hat{\mathbf{y}} \leftarrow \arg\max(\hat{\mathbf{p}})$
  **Compute** uncertainty: confidence $= \max \hat{\mathbf{p}}$, entropy $= H(\mathbf{x})$
  **Compute** NLL, Brier score, ECE and all classification metrics; save per-split results.
**end for**
Aggregate metrics over splits; generate calibration and diagnostic plots.

### V.3.3 Results: Skeletal Muscle Channelopathies Dataset

#### Quantitative comparison across individual models

Across the four candidate models and the two wavelet representations the quantitative picture is remarkably coherent (Tab. V.5). Validation balanced-accuracy clusters tightly between 77 % and 85 %, with Inception-ResNetV2 on Morse scalograms sitting at the top (84.4 %), and MaxViT on Morlet at the bottom (77.3 %). The dispersion on the independent test sets is even smaller, confirming that no single architecture catastrophically fails. Because of this homogeneity it is sufficient to dissect in detail the behavior of the InceptionResNetV2 on Morse transformed dataset configuration, which will serve as the reference in the remainder of the subsection.

When predictions from all five cross-validation folds are pooled, the reference network attains an area under the ROC curve of 0.90 on validation and 0.84 on test set (Fig. V.22). The corresponding average-precision values are 0.93 and 0.87 indicating that the classifier preserves a favorable precision-recall trade-off even under the class skew. Accuracy and balanced-accuracy on the validation splits converge at 84.4 %, while they drop to approximately 75 % on the unseen test files.

The class-wise analysis of Fig. V.22a and Fig. V.22d reveals that chloride myotonias (Cl) are recognized more readily than sodium channelopathies (Na). On the test ensemble the true-positive rate for chloride-channel defect class reaches 79.4 %, whereas only 70.1 % of sodium-channel defect samples are correctly labeled. Missed Na cases therefore form the dominant error mode (47 samples, 29.9 % of all sodium cases), a pattern that re-appears across individual folds and architectures.

Among the five cross-validation folds, split 5 is systematically the weakest, whereas the top scores are always attained on either split 3 or split 4. InceptionResNetV2 peaks on split 3 when trained with Morse wavelet transformed scalograms, but on split 4 with Morlet scalograms. Conversely, MaxViT performs best on split 4 for Morse and on split 3 for Morlet, confirming that the "optimal" fold depends as much on patient composition as on backbone or wavelet.

The three folds are almost identical in size – split 3: 403/53/44, split 4: 378/67/55, split 5: 382/53/65 files for train/validation/test, respectively, so the performance gap cannot be attributed to sample count alone. During the selection process, we noticed that some files had better quality with pronounced myotonic discharges, sometimes several on the recording, while others had barely distinguishable single discharge and heavy contamination. These might be the factors that influenced classification results.

The fold-to-fold comparison in Tab. V.6 and Fig. V.23 illustrates how sensitive performance remains to the exact composition of the patient groups. Unlike split 5, splits 3 and 4 share two patient samples (IDs 37 and 22) which might be the samples with more pronounced discharges and thus easier to classify. This could explain why InceptionResNetV2 model reaches ROC-AUC of 0.91 and balanced accuracy of 80.4 % on the test cohort of split 3, yet loses 11-13 percentage points on every metric when evaluated on split 5. This drop is consistent with the marked inter-patient heterogeneity that myotonic EMG data typically exhibit. Taken together, the evidence supports the view that clinical phenotype, discharge morphology and recording quality, rather than model instability or data volume exclusively drive the observed variance across folds.

Taken together, the proposed deep-learning models discriminate between chloride and

sodium channelopathies considerably better than chance and maintains high precision under realistic class imbalance (AP of around 0.87). The residual error is not evenly distributed but concentrates on sodium cases. This finding is clinically relevant: the algorithm is less prone to overlook chloride-channel defect myotonia, which carry greater diagnostic consequences.

**Table V.5:** Averaged Metrics over Five Splits for Individual Classifiers

| Model Wavelet | Set | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Incept. Morse | Val. | 84.4 ± 0.084 | 84.4 ± 0.084 | 84.4 ± 0.084 | 84.5 ± 0.080 | 84.3 ± 0.083 | 0.90 ± 0.050 | 92.6 ± 0.050 | 0.69 ± 0.170 | 84.4 ± 0.084 | 0.79 ± 0.048 | 0.48 ± 0.056 |
| | Test | 74.7 ± 0.036 | 74.8 ± 0.035 | 75.5 ± 0.033 | 75.5 ± 0.033 | 74.2 ± 0.037 | 0.85 ± 0.037 | 88.0 ± 0.044 | 0.48 ± 0.076 | 74.5 ± 0.038 | 0.78 ± 0.040 | 0.48 ± 0.052 |
| Incept. Morlet | Val. | 80.7 ± 0.051 | 80.7 ± 0.051 | 80.7 ± 0.051 | 81.1 ± 0.053 | 80.6 ± 0.053 | 0.87 ± 0.051 | 89.1 ± 0.055 | 0.61 ± 0.105 | 80.7 ± 0.051 | 0.77 ± 0.046 | 0.50 ± 0.060 |
| | Test | 74.0 ± 0.065 | 74.4 ± 0.059 | 74.5 ± 0.061 | 74.5 ± 0.061 | 73.7 ± 0.065 | 0.83 ± 0.067 | 86.2 ± 0.044 | 0.52 ± 0.108 | 76.2 ± 0.051 | 0.76 ± 0.061 | 0.51 ± 0.074 |
| MaxViT Morse | Val. | 80.1 ± 0.051 | 80.1 ± 0.051 | 80.1 ± 0.051 | 80.1 ± 0.050 | 80.0 ± 0.050 | 0.88 ± 0.044 | 90.6 ± 0.034 | 0.60 ± 0.096 | 80.2 ± 0.049 | 0.76 ± 0.038 | 0.51 ± 0.041 |
| | Test | 74.8 ± 0.076 | 75.4 ± 0.069 | 76.5 ± 0.066 | 76.5 ± 0.066 | 74.3 ± 0.077 | 0.82 ± 0.090 | 86.8 ± 0.048 | 0.50 ± 0.145 | 75.4 ± 0.069 | 0.76 ± 0.023 | 0.51 ± 0.026 |
| MaxViT Morlet | Val. | 77.3 ± 0.093 | 77.3 ± 0.093 | 77.3 ± 0.093 | 77.7 ± 0.092 | 77.1 ± 0.094 | 0.87 ± 0.059 | 89.9 ± 0.036 | 0.55 ± 0.186 | 77.3 ± 0.093 | 0.76 ± 0.051 | 0.50 ± 0.059 |
| | Test | 73.7 ± 0.083 | 74.2 ± 0.072 | 75.7 ± 0.070 | 75.7 ± 0.070 | 73.1 ± 0.086 | 0.81 ± 0.073 | 85.9 ± 0.044 | 0.48 ± 0.152 | 74.2 ± 0.072 | 0.76 ± 0.045 | 0.51 ± 0.059 |

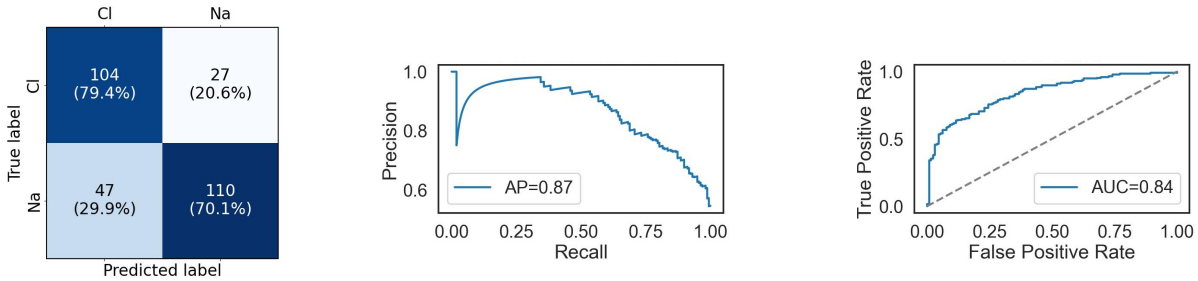
**Note:** Rec. = Recall; Spec. = Specificity; AP = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Confidence; Ent. = Entropy; Val. = Validation; Incept. = InceptionResNetV2. All metrics are reported as mean ± standard deviation across the five splits.

To improve the results one should focus on mitigating the fold-to-fold variability. Enlarging the training corpus with targeted augmentations of low-energy, noise-dominated samples and incorporating a signal-quality weighting term in the loss function are promising directions to enhance performance on the hardest sodium cases while keeping the current robustness on chloride recordings.

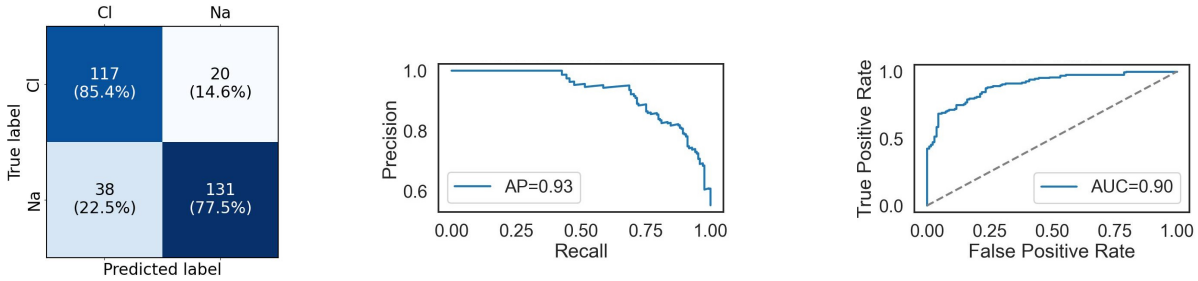
**Table V.6:** Metrics Summary for InceptionResNetV2 (Morse) on Splits 3 and 5

| Set | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Split 3 Val. | 90.6 | 90.6 | 90.6 | 90.0 | 90.2 | 0.96 | 97.5 | 0.80 | 90.6 | 0.82 | 0.44 |
| Split 3 Test | 79.6 | 80.4 | 80.7 | 80.7 | 79.5 | 0.91 | 93.1 | 0.60 | 80.4 | 0.82 | 0.45 |
| Split 5 Val. | 77.3 | 77.3 | 77.3 | 77.3 | 77.3 | 0.87 | 86.3 | 0.55 | 77.3 | 0.79 | 0.48 |
| Split 5 Test | 69.2 | 71.2 | 70.6 | 70.6 | 69.2 | 0.80 | 85.1 | 0.40 | 71.2 | 0.82 | 0.43 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; AP = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Confidence; Ent. = Entropy; Val. = Validation.

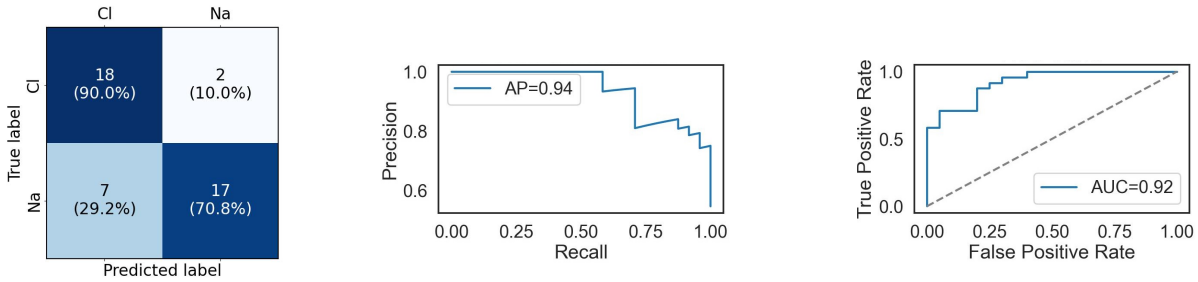



**(a)** Confusion matrix **(b)** Precision-recall curve **(c)** ROC curve

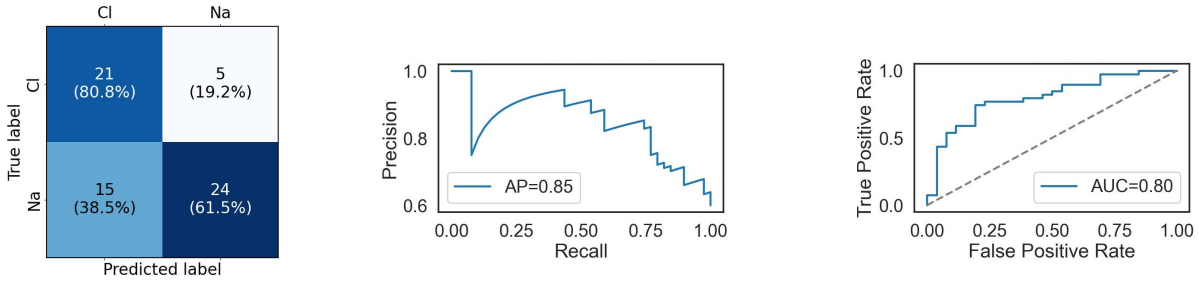



**(d)** Confusion matrix **(e)** Precision-recall curve **(f)** ROC curve

**Figure V.22:** Aggregated classification performance of the InceptionResNetV2 model on Morse wavelet transformed dataset averaged across all splits on the **test set (top row)** and the **validation set (bottom row)**.




**(a)** Confusion matrix **(b)** Precision–Recall curve **(c)** ROC curve




**(d)** Confusion matrix **(e)** Precision–Recall curve **(f)** ROC curve

**Figure V.23:** Test-set performance for InceprtionResNetV2 classifier on Morse wavelet transformed dataset for **Split 3 (top row)** and **Split 5 (bottom row)**.

**Gradient-weighted Class Activation Mapping Analysis of Model Decisions**

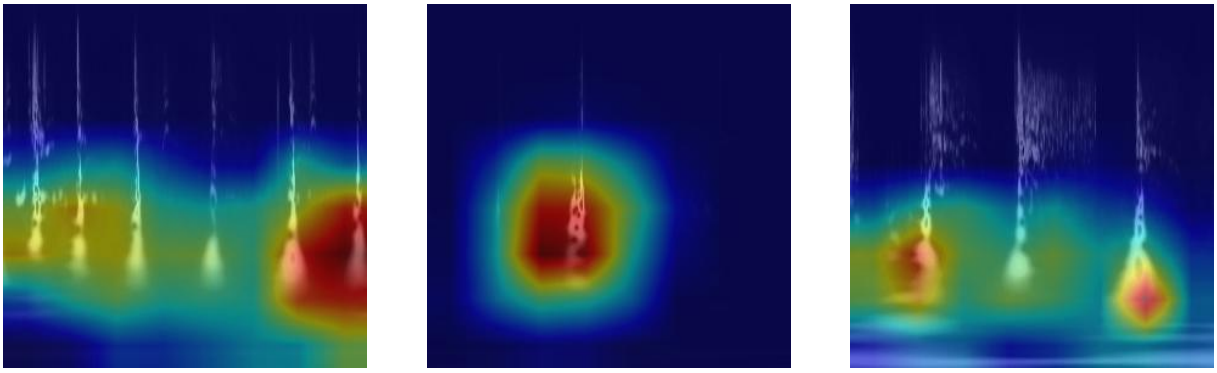
Gradient-weighted Class Activation Mapping (Grad-CAM) back-propagates the gradient $\frac{\partial y_c}{\partial A^k_{ij}}$ of the class score $y_c$ with respect to the feature-map $A^k$ of a chosen convolutional layer, averages that gradient over the spatial indices $i, j$ to obtain a scalar weight $\alpha^c_k$, and then forms a weighted sum $\mathrm{ReLU}\big(\sum_k \alpha^c_k A^k\big)$ (Selvaraju et al., 2017). The result is an image-sized heat map whose pixel intensity is proportional to the positive influence that the corresponding receptive field exerts on the class score. Because our inputs are scalograms, every pixel has a well-defined meaning in *time* and *frequency*. A warm pixel therefore states, in millisecond-hertz coordinates, where the network has found class-defining evidence. Unlike occlusion tests, the method is fast (one backward pass) and produces a single, smooth map that can be over-laid on the original scalogram without changing its geometry.

Grad-CAM was applied to the last convolutional layer of the trained network at inference stage. Each map was min-max normalized, ReLU-clipped, and blended with the grayscale scalogram at 30 % opacity; the processing pipeline was identical for correctly and incorrectly classified traces, so any systematic difference in saliency must arise from the data rather than from the visualization. We analyzed Grad-CAM of test set samples to find out about the image parts which have influenced the classification results most.

Drost et al. (2015) reported that the first inter-discharge interval (IDI) was longer than 30 ms in 31/32 chloride-channel cases and shorter than 30 ms in 34/34 sodium-channel cases, making the 30-ms threshold a nearly perfect separator for rectus femoris muscle. In frequency terms the criterion corresponds to a fundamental firing rate of $f_0 = 1000/\mathrm{IDI} \approx 33\,\mathrm{Hz}$. Because a single motor-unit potential is broadband, its harmonics rise well above 1 kHz, spikes that recur faster than 30 ms produce overlapping high-frequency tails; on the scalogram this appears as a *continuous horizontal band* of power. When the interval is longer, high-frequency power has time to decay, so each spike appears as an *isolated vertical column* that is separated from its neighbor by a cold (low-power) gap. Thus the Grad-CAM patterns observed in our data – one uninterrupted band for sodium recordings and several discrete streaks for chloride recordings could reflect the temporal dichotomy quantified by Drost et al. (2015) in rectus femoris muscle.

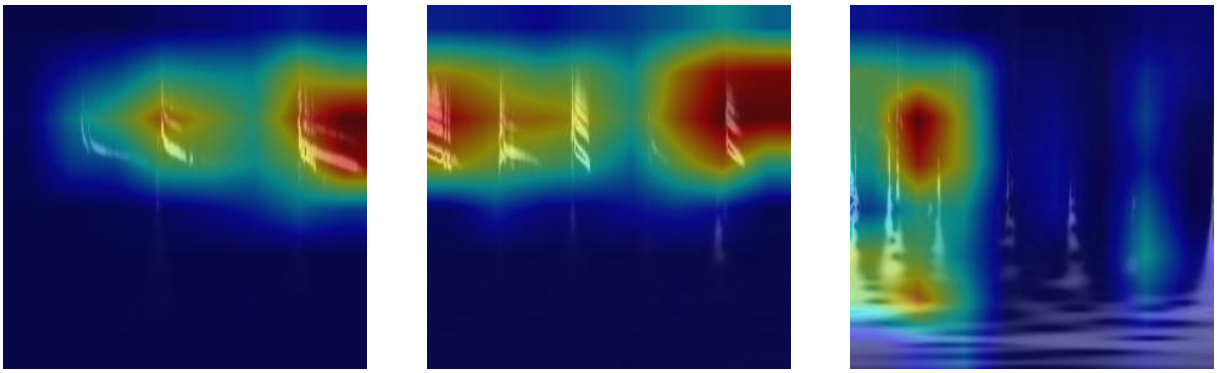
In the three chloride examples, depicted in Fig. V.24 (`p10 CPR`, `p37 CR`, `p49 CPR`) the heat map breaks into several individual blobs, one per myotonic discharge. Each blob sits on the low-to-mid portion of the broadband column (roughly 100–600 Hz); the network assigns little weight to the very low-frequency foot, where absolute energy is highest, and ignores the extreme apex, where the signal approaches the noise floor. The gaps between hot blobs remain mostly cold, confirming that high-frequency power falls to baseline before the next discharge begins. Separate blobs therefore might indicate an intra-discharge IDI longer than 30 ms consistent with the chloride profile reported by Drost et al. (2015).

All sodium examples (p55 SD, p16 SD, p39 SD) display a *single contiguous patch* that spans time and frequency. the Grad-CAM heat concentrates in the high-frequency tail (around 800–1200 Hz) of each dense burst. The warm region usually does not drop to baseline; the uninterrupted high frequency power might indicate IDI $< 30$ ms and thus consistent with finding of Drost et al. (2015). The differences between class samples are shown in more detail in Fig. V.26.



**(a)** p10 CPR **(b)** p37 CR **(c)** p49 CPR

**Figure V.24:** Grad-CAM overlays for three correctly predicted chloride-channel myotonia recordings. Warm colors mark the broadband apex of each burst, while the cold gaps between bursts indicate complete decay of high-frequency power.

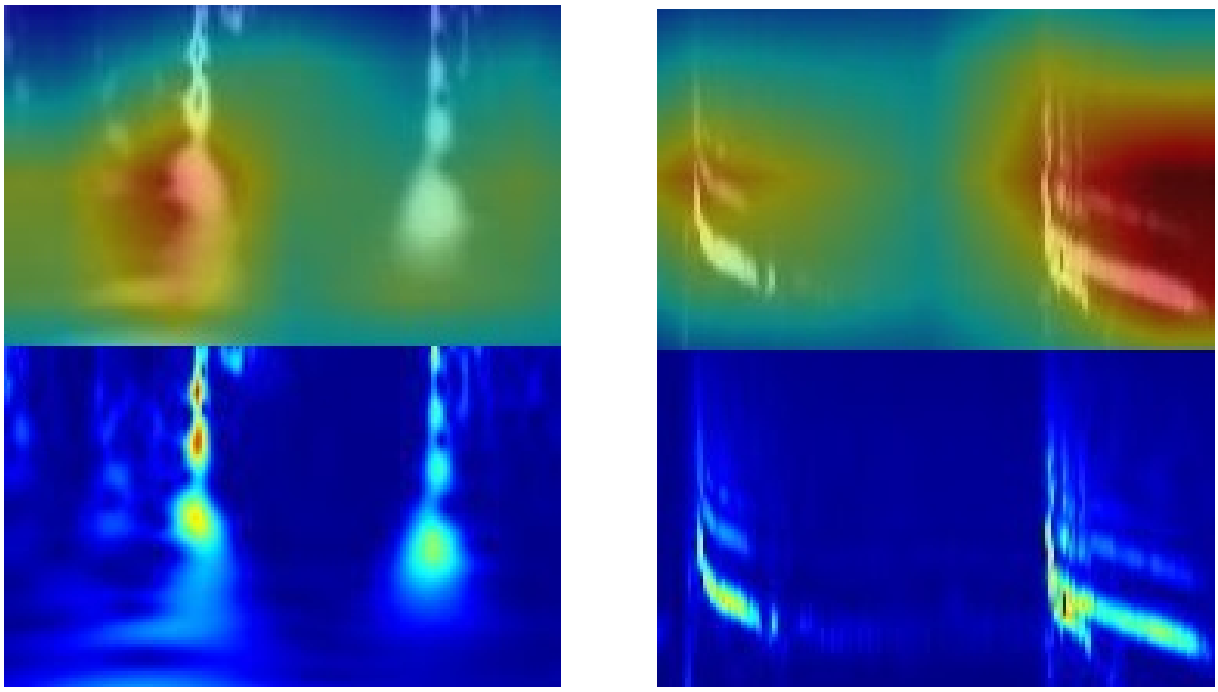


**(a)** p55 SD **(b)** p16 SD **(c)** p39 SD

**Figure V.25:** Grad-CAM overlays for three correctly predicted sodium-channel myotonia recordings. In contrast to the chloride examples, saliency coalesces into a single continuous patch that spans the high-frequency band throughout the dominant burst, indicating uninterrupted power.

Across the entire measured dataset Grad-CAM implies that the CNN detects physiologically meaningful discriminants. Multiple broadband spikes separated by high-frequency silence are taken as evidence for chloride myotonia, whereas a single, uninterrupted high-band ridge is taken as evidence for sodium myotonia. The only systematic failure mode occurs

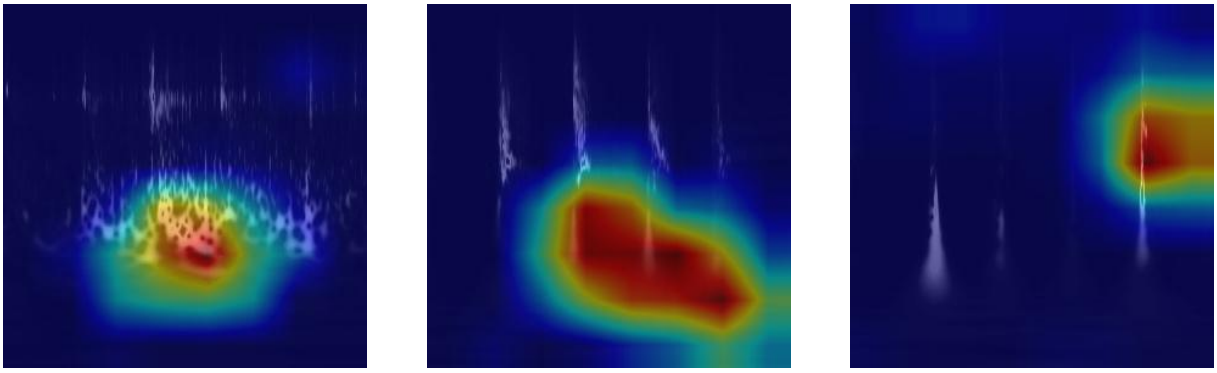
when a single burst dominates the trace's amplitude profile, suppressing weaker but diagnostically relevant context.



**(a)** Chloride example (`p49 CPR`) **(b)** Sodium example (`p55 SD`)

**Figure V.26:** Zoomed-in composites (**upper half**: Grad-CAM, **lower half**: scalogram of the image's same region) highlight the class-specific microstructure used by the network. **(a) Chloride.** Grad-CAM concentrates on the low-to-mid frequency portion of vertical streak. The network gives greater weight to spectral *breadth* despite lower absolute power in that band. **(b) Sodium.** Power is confined to a narrow, oblique ridge. Grad-CAM aligns with this high-band ridge, indicating that the classifier relies on *continuity in a restricted band*. Together the zooms demonstrate that the CNN distinguishes chloride from sodium myotonia by bandwidth and persistence, not by raw signal magnitude.

On three misclassified samples depicted in Fig. V.27, it can be seen that the attention map collapsed onto that dominant segment and ignored the contextual pattern: burst count for chloride and continuity for sodium myotonia class. When the loud burst mimicked the morphology of the opposite class the prediction flipped; when it retained class-typical shape the label remained correct. Amplitude, therefore, can act not as an alternative cue but as a gate that can hide the learned physiological features. Misclassifications arise only when one burst dominates the dynamic range and masks weaker context, suggesting that a simple dynamic-range normalization or a context-aware attention mechanism could eliminate the remaining errors.



**(a)** True Na, predicted Cl **(b)** True Na, predicted Cl **(c)** True Cl, predicted Na

**Figure V.27:** Grad-CAM overlays for the three retained misclassified recordings. **Left and center**: sodium-channel myotonias wrongly labeled chloride: saliency collapses on the loudest broadband segment, suppressing the band-limited ridge that normally identifies sodium pathology. **Right**: a chloride-channel myotonia mislabeled sodium: one dominant burst attracts all the attention, so the network overlooks earlier broadband bursts and reproduces the single-patch pattern typical of sodium cases. The figure illustrates the model's sensitivity to extreme intra-trace amplitude imbalance.

### Ensemble Performance

As the next step, we ensembled the obtained models to get their predictions on the test sets for each split, which will be our ultimate test for the developed algorithms. The Tab. V.7 shows the aggregated averaged performance of the models over all five splits.

The four-network ensemble produced consistently strong results across the five cross-validation splits. On the test partitions it achieved a mean ROC-AUC of $0.89 \pm 0.06$ and an average-precision of $0.92 \pm 0.02$, together with a balanced accuracy of $0.81 \pm 0.04$ (Tab. V.7). These values exceed the average test performance of the twenty single models by 5-7 percentage points, confirming that the heterogeneous mixture of two InceptionResNetV2 and two MaxViT backbones successfully reduces both bias and variance.

The aggregated confusion matrix (Fig. V.28a) reveals an asymmetric error pattern, which was also present in single model performances: chloride myotonia is detected with a recall of 90 % whereas sodium channelopathy attains 72 %. In absolute terms, 44 sodium recordings are mislabeled as chloride, compared with only 13 errors in the opposite direction. This residual bias most likely reflects the larger intra-class variability observed in sodium samples and the modest degree of class imbalance ($\sim$ 45 % Cl and 55 % Na) rather than systematic overfitting, because the model never collapses on any fold.

Probabilistic outputs are well behaved. The reliability curve (Fig. V.28d) tracks the

diagonal closely; the expected calibration error is $0.11 \pm 0.05$ and the Brier score $0.14 \pm 0.03$. Selective prediction analysis (Fig. V.28e) further shows that discarding samples whose confidence falls below 0.85 would raise accuracy to $\approx 0.90$.However,this higher accuracy comes at the cost of potentially making predictions on fewer cases. The graph strongly advises against setting a threshold above approximately 0.9, as this would lead to a dramatic loss of accuracy, rendering the model's high-confidence predictions unreliable for decision support. The predictive-entropy histogram (Fig. V.28f) corroborates this showing that for a considerable number of predictions, the model exhibits high uncertainty. This means it frequently does not make very "strong" predictions where one class probability dominates overwhelmingly.

The strongest ensemble performance is registered for split 3 (Tab. V.8), as indicated by excellent ROC-AUC (0.95) and Average Precision (0.97) scores. There is a notable expected calibration error (ECE of 0.14), suggesting the predicted probabilities are not perfectly calibrated to reflect true accuracy. Nevertheless, the high Cohen's Kappa of 0.73 indicates substantial agreement, confirming performance significantly better than random chance. The Fig. V.29 suggests that the model is slightly under-confident in its higher probability predictions, meaning its accuracy in those confidence bins is higher than its predicted confidence. While the model often produces predictions with moderate to high uncertainty, very high accuracy can be achieved by only considering its most confident predictions.

**Table V.7:** Aggregated Metrics for Ensemble Predictions across Splits

| Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. | NLL | Brier Score | ECE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80.7 ± 0.049 | 81.4 ± 0.042 | 81.4 ± 0.042 | 81.7 ± 0.040 | 80.5 ± 0.050 | 0.89 ± 0.055 | 92.4 ± 0.023 | 0.62 ± 0.091 | 81.4 ± 0.042 | 0.74 ± 0.019 | 0.54 ± 0.022 | 0.46 ± 0.072 | 0.14 ± 0.027 | 0.11 ± 0.045 |

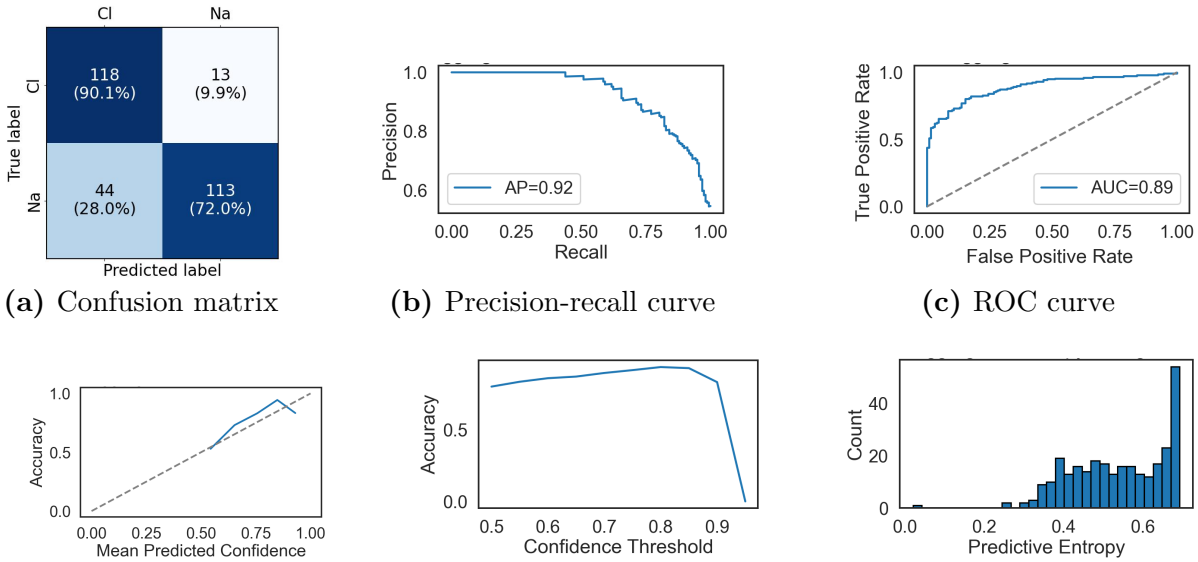
**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1. = F1 Score; AP = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Mean Confidence; Ent. = Mean Entropy. All metrics are reported as mean ± standard deviation across the five splits.

**Table V.8:** Metrics for Ensemble Predictions for Selected Splits

| Split # | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. | NLL | Brier Score | ECE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 86.4 | 86.7 | 86.7 | 86.4 | 86.3 | 0.95 | 96.6 | 0.73 | 86.7 | 0.74 | 0.53 | 0.37 | 0.11 | 0.14 |
| 5 | 72.3 | 75.0 | 75.0 | 74.7 | 72.3 | 0.81 | 89.5 | 0.46 | 75.0 | 0.75 | 0.53 | 0.57 | 0.19 | 0.04 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1. = F1 Score; AP = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Mean Confidence; Ent. = Mean Entropy.

Split 5 constitutes the lower bound of ensemble's discriminative performance. It attains an overall accuracy of 72.3 %, a ROC-AUC of 0.81, and a Cohen's Kappa of 0.46,




**(a)** Confusion matrix **(b)** Precision-recall curve **(c)** ROC curve

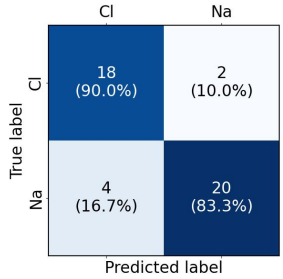
**(d)** Reliability curve **(e)** Selective-prediction curve **(f)** Distribution of entropy

**Figure V.28:** Aggregated results for the ensembled classifier across all test splits. **Top row**: class-wise errors and discrimination ability. **Bottom row**: reliability and usefulness of the model's probabilistic outputs. These plots complement the metrics reported in Tab. V.7.

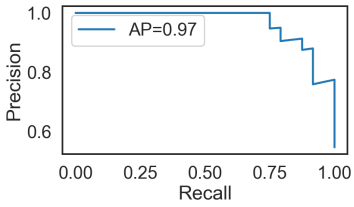
all noticeably below the figures for Split 3. The class-wise results are imbalanced: chloride-channel myotonia is recovered with a recall of 88.5 %, whereas sodium-channel myotonia reaches only 61.5 %. On the positive side, the model's probabilistic outputs are highly trustworthy: its expected calibration error is just 0.04, indicating that the predicted probabilities closely match the true outcome frequencies.

Fig. V.31 summarizes the ensemble's performance across five different data splits by showing the raw count of correct versus incorrect classifications. Overall, the model consistently makes a significantly higher number of correct predictions than incorrect ones across all splits, indicating a generally good performance. However, the variability is also present; while splits 1 and 2 show the highest number of correct classifications and relatively few errors, split 5 exhibits a notable increase in incorrect predictions, suggesting it might be a more challenging subset for the model.

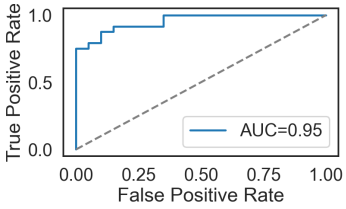
The ensemble attains a level of discrimination that approaches clinical utility while providing reasonably calibrated uncertainty estimates. Most residual errors stem from sodium traces with atypical temporal morphology or reduced signal-to-noise ratio, suggesting that the current feature set still lacks descriptors sensitive to those patterns. Future work could focus on augmenting sodium examples and applying post-hoc temperature scaling or deep-ensemble distillation to tighten calibration without sacrificing accuracy.



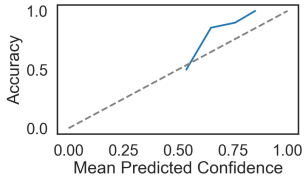

(a) Confusion matrix



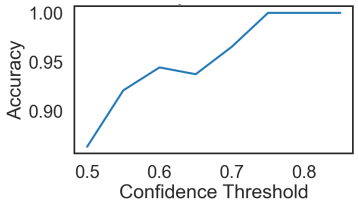

(b) Precision–recall curve



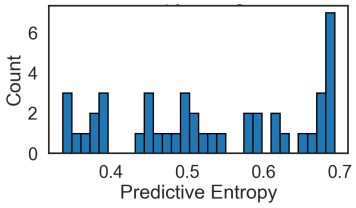

(c) ROC curve




(d) Reliability curve

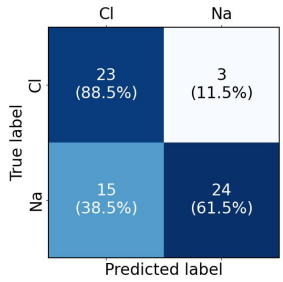



(e) Selective-prediction curve

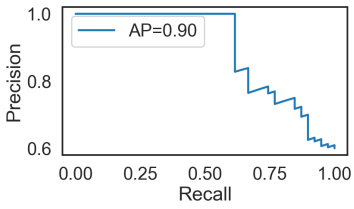



(f) Distribution of entropy

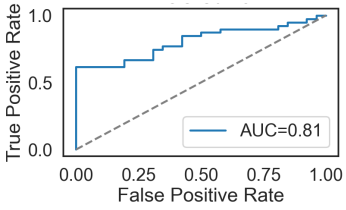
**Figure V.29:** Split 3 results for the ensembled classifier. **Top row**: class-wise errors and discrimination ability. **Bottom row**: reliability and usefulness of the model's probabilistic outputs. These plots complement the metrics reported in Tab. V.8.



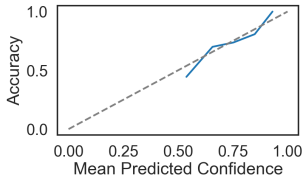

(a) Confusion matrix



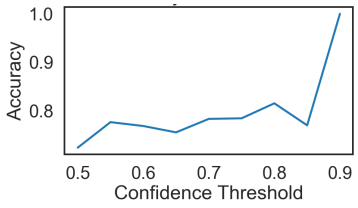

(b) Precision–Recall curve



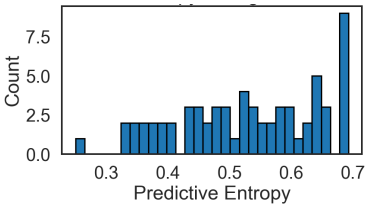

(c) ROC curve




(d) Reliability curve




(e) Selective-prediction curve




(f) Distribution of entropy

**Figure V.30:** Split 5 results for the ensembled classifier. **Top row**: class-wise errors and discrimination ability. **Bottom row**: reliability and usefulness of the model's probabilistic outputs. These plots complement the metrics reported in Tab. V.8.

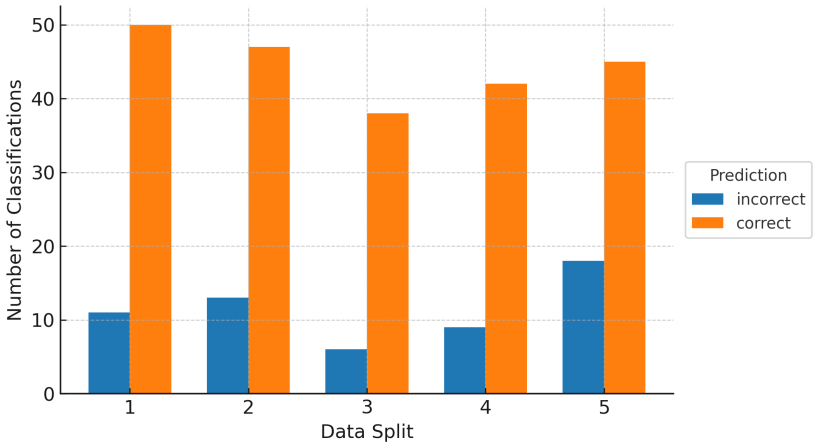



**Figure V.31:** Bar plot showing the number of correct (orange) and incorrect (blue) classifications for each of the five data splits in our experiment. This highlights how model performance varies across different test partitions.

### V.3.4 Initial Single-Model Test: Simulated Skeletal Muscle Channelopathy Dataset

Single ImageNet model classification pipeline (without data augmentation) was initially tested on the simulated muscle channelopathy dataset. This allowed us to assess the proposed method under controlled conditions, as described in Subsection III.1.2, and to incorporate a healthy control group. Our underlying assumption was that if the method successfully discriminates classes within synthetic data, where discriminatory features are intentionally embedded, it would more likely generalize effectively to measured clinical data. Although a human evaluator can typically distinguish simulated healthy iEMG signals from those containing myotonic discharges, differentiating specific channel defect types visually from amplitude-time waveforms alone remains impossible. Thus, this scenario explicitly demonstrates the existence of hidden discriminative patterns in the data, invisible to human observers but identifiable by the classification algorithm.

Across the five cross-validated splits the single-model pipeline based on InceptionResNetV2 reached virtually perfect discrimination on the simulated channelopathy dataset. All aggregate metrics in Tab. V.9 remained at or above 0.999 on both the validation and test sets, and the 95 % confidence intervals collapsed to the fourth decimal place, indicating an absence of performance variability. Mean predictive confidence was very high ($\approx$0.86) while the mean predictive entropy stayed below 0.50, showing that the network expressed a strong preference for a single class in almost every instance. Out of 2172 test instances only three images were mislabeled, all originating from subject 34 with a simulated sodium-channel defect in split 4, and all erroneously assigned to the healthy class.

**Table V.9:** Aggregated Validation and Test Metrics

| Set | Acc. | Rec. | Spec. | Prec. | F1 | $\kappa$ | Conf. | Ent. |
|---|---|---|---|---|---|---|---|---|
| Val. | 100.00 ±0.0000 | 100.00 ±0.0000 | 100.00 ±0.0000 | 100.00 ±0.0000 | 100.00 ±0.0000 | 1.0000 ±0.0000 | 0.8624 ±0.0029 | 0.4955 ±0.0072 |
| Test | 99.95 ±0.0009 | 99.95 ±0.0009 | 99.95 ±0.0009 | 99.95 ±0.0009 | 99.95 ±0.0009 | 0.9993 ±0.0014 | 0.8608 ±0.0046 | 0.4985 ±0.0118 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1 = F1 score; $\kappa$ = Cohen's kappa; Conf. = Confidence; Ent. = Entropy; Val. = Validation. All metrics are reported as mean ± standard deviation across the five splits.

The biophysical model of Klotz et al. (2020) produces distinct, deterministic patterns in the chloride- and sodium-channel defect myotonic discharges, such that frequency-time signatures of synthetic EMG scalograms whose cluster tightly inside class-specific manifolds. This reduces intra-class variance and yields large inter-class margins that a deep neural network can exploit easily. Unlike clinical raw data, the simulations are free of motion artifacts and spontaneous background activity, eliminating confounders that would normally blur class boundaries.

The initial single-model experiment establishes that an ImageNet-pretrained InceptionResNetV2 can achieve near-perfect tri-class separation on noise-free, biophysically faithful simulations of chloride and sodium channelopathies. The minuscule error count and the introspective uncertainty estimates highlight the network's capacity to recognize its own decision boundaries.

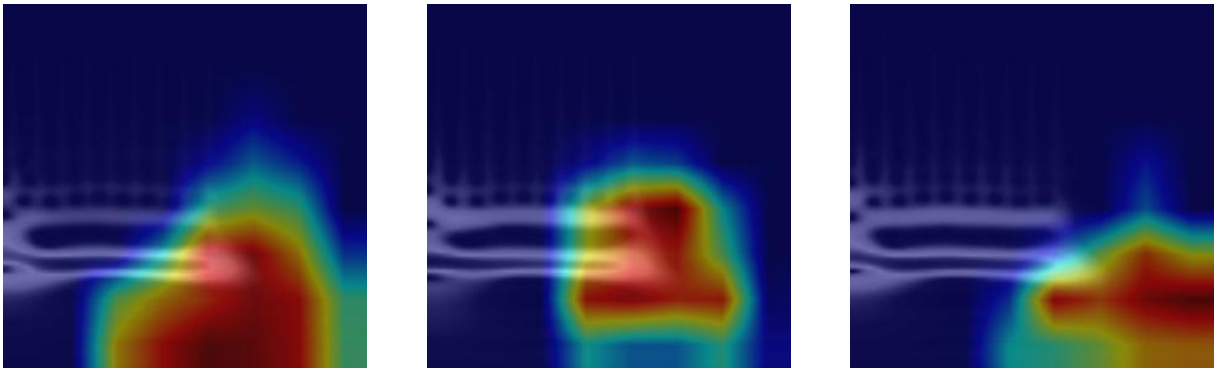
### Gradient-weighted Class Activation Mapping Analysis of Model Decisions

The network that attained almost perfect quantitative performance on the simulated test set exhibits an equally coherent qualitative behavior. Visual inspection of the Grad-CAM overlays (Fig. V.33-V.34) shows that in every correctly classified trial the network attention concentrates on the segment that carries the myotonic discharge, while the remainder of the time-frequency plane is ignored. Because this spatial focus coincides with the envelope of motor-unit activity rather than with background regions devoid of biological content, the visual evidence corroborates that the classifier exploits genuine signal morphology.

Although all three diagnostic categories share this common locus of attention, their saliency patterns are characteristically different. Healthy recordings are marked by a compact, right-shifted activation patch that fades rapidly after the terminal burst (Fig. V.32). This pattern shows that the network the network bases its decision primarily on the absence of post-stimulus activity as the decisive hallmark of normal physiology. For chloride-channel dysfunction the Grad-CAM highlights a teardrop-shaped region that starts just after the ninth stimulus and extends $\approx 20 - 30$ ms into the discharge epoch (Fig. V.33). The

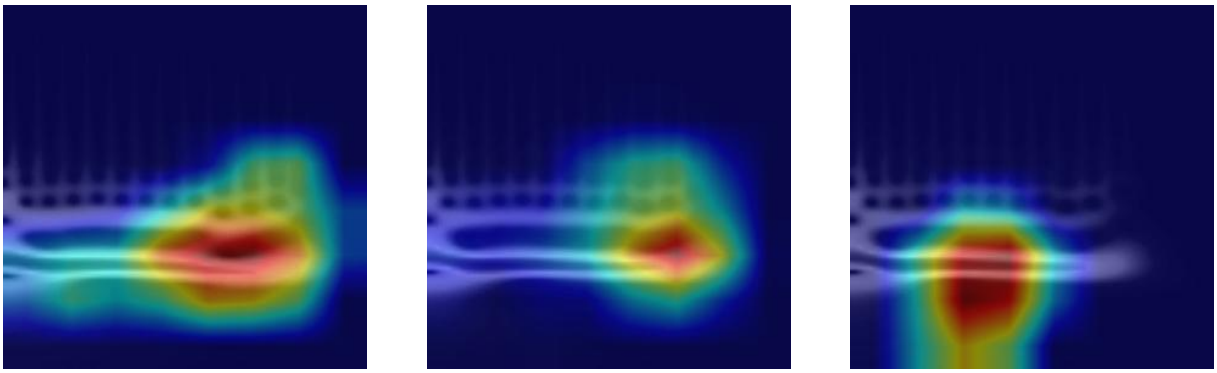
limited lateral spread along time-axis is consistent with a burst that fades comparatively quickly; the slight vertical elongation reflects the moderate frequency content (200–500 Hz).

The sodium class maps differ in two systematic ways (Fig. V.34). First, the salient zone occupies a much longer segment along the horizontal axis, mirroring the prolonged, waxing–waning firing that continues well beyond the window captured in chloride traces. Second, the activation band is flatter and biased towards lower frequencies, in line with the larger depolarizing drive and the consequent recruitment of more slowly conducting fibers. Thus, the classifier distinguishes sodium from chloride channelopathy chiefly through the longer duration and flatter spectral footprint of the post-stimulus discharge, while also re-weighing the immediately preceding stimulus-locked spikes that frame this activity.



**(a)** Split 3 **(b)** Split 4 **(c)** Split 5

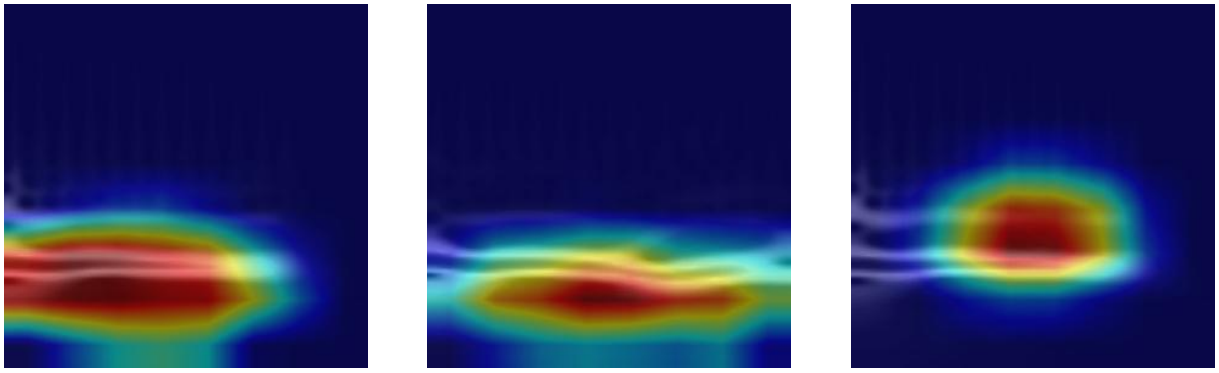
**Figure V.32:** Grad-CAM saliency maps for healthy controls. In these examples, attribution concentrates late in the window with limited temporal spread.



**(a)** Split 3 sample **(b)** Split 4 sample **(c)** Split 5 sample

**Figure V.33:** Grad-CAM saliency maps for chloride-channel myotonia across three splits. The highlighted region is typically shorter in time than in many sodium-channel examples.

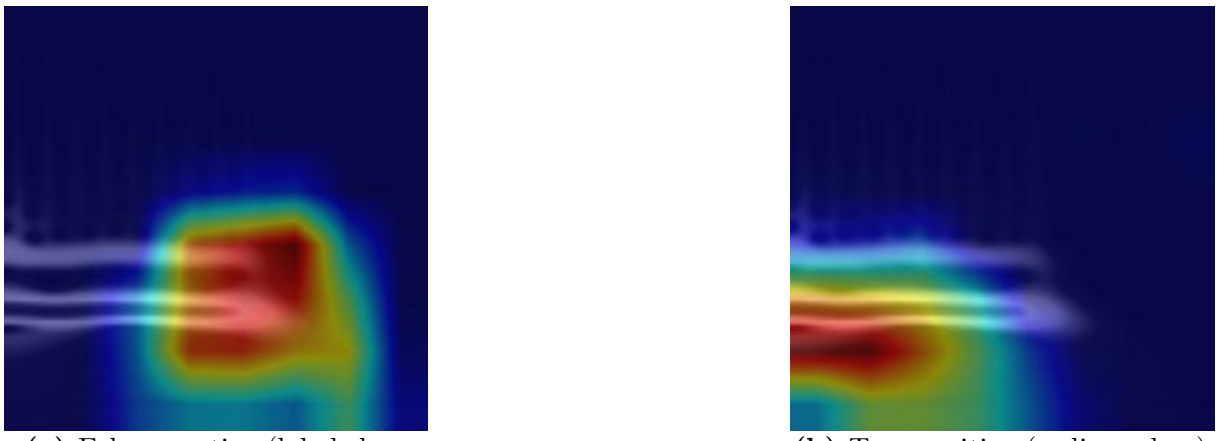
Only in three cases the network misclassified sodium-channel myotonia samples as healthy. In all of these three false negatives, discharges petered out unusually early, leaving only a



**(a)** Split 3 sample **(b)** Split 4 sample **(c)** Split 5 sample

**Figure V.34:** Grad-CAM saliency maps for sodium-channel myotonia. Many examples show a horizontally extended attribution band relative to chloride-channel cases.

short after-potential that mimicked the healthy template, which can be seen in Fig. V.35 comparing incorrectly classified sodium class sample to correctly classified one. Their Grad-CAM overlays shrink to the compact footprint typical of the healthy class. Apart from this isolated scenario the decision boundaries between the three classes remain clean, an observation that is consistent with the quantitative metrics.



**(a)** False negative (labeled as healthy **(b)** True positive (sodium class)

**Figure V.35:** Representative Grad-CAM overlays for patient 34 (sodium-channel myotonia). **(a)** The misclassified trial shows a compact, right-shifted activation block that mimics the healthy template. **(b)** The correctly classified trial, recorded from a different simulated needle position, exhibits the canonical horizontally elongated sodium class pattern. The two additional false-negative trials from this patient display near-identical saliency geometry.

Although the CNN achieved near-perfect accuracy on the synthetic test set, Grad-CAM analysis reveals that saliency patterns for both simulated myotonia classes predominantly activate frequency components below approximately 500 Hz. This implies that the model

relies exclusively on low-frequency information, since all diagnostic features required for discriminating the classes are embedded within this limited spectral range. Clinical EMG signals, however, contain broadband activity and high-frequency spikes extending into the kilohertz range. Consequently, the network's current spectral filters, trained solely on synthetic data lacking such features, remain unprepared for real-world recordings. Thus future simulations should incorporate realistic broadband and high-frequency spike content to ensure practical clinical application of synthetically-trained classifiers or exploitation of synthetic data for data augmentation purposes.

### V.3.5 Results: Fibrillation Potentials Dataset

#### Single Model Performance

The quantitative evaluation of both InceptionResNetV2 and MaxViT architectures on the Morse- and Morlet-wavelet transformed scalograms demonstrates that the task of differentiating pathological fibrillation potentials (Fib) from voluntary contraction potentials (Vol) can be solved with high, yet still imperfect, reliability. Tab. V.10 shows that MaxViT trained on the Morse scalograms delivers the best aggregate performance, reaching on the test folds a balanced accuracy of 86.2 %, and an ROC-AUC of 0.94. InceptionResNetV2 on the same representation lags behind by roughly five percentage points, while both networks suffer a modest ($< 3\,\%$) drop when the Morlet wavelet is used, indicating that the Morse wavelet captures shape descriptors that are more discriminative for the sharp, short-duration Fib bursts.

The slightly lower mean entropy (Tab. V.10) for Morse-based models compared to Morlet-based models may indicate a subtly more defined feature space captured by Morse wavelets. This might translate into slightly more confident predictions, even though the overall confidence scores remain quite similar between models. The consistency between validation and test metrics, with differences mostly within one standard deviation, suggests strong generalizability and minimal overfitting. The single notable exception was MaxViT on split 3, where an unexpectedly higher accuracy was recorded on the test set (92 %) compared to the validation set (74 %). This anomaly may be explained by the test subset being coincidentally "easier" for this particular network configuration, as all other splits showed the expected minor performance decline when evaluating unseen data.

Focusing on the aggregated test predictions for the InceptionResNetV2 model using Morse scalograms, as illustrated in Figure V.36 (top row), the confusion matrix reveals a relatively balanced error profile. The model achieved a sensitivity (recall for Fib) of 77.4 % and a specificity (recall for Vol) of 86.1 %. The strong threshold-independent metrics, namely an average precision of 0.88 and a ROC-AUC of 0.89, further corroborate the substantial separability between classes. Cohen's kappa of approximately 0.60 signifies a substantial level of agreement between the model's predictions and the true labels, indicating a performance well above chance. The comparison with the validation set

performance (Figure V.36, bottom row) shows broadly similar patterns, with validation ROC-AUC at 0.88 and AP at 0.91, reinforcing the generalisability observed in the averaged metrics table.

**Table V.10:** Averaged Metrics over Splits for InceptionResNetV2 and MaxViT Classification of Fibrillation Potentials

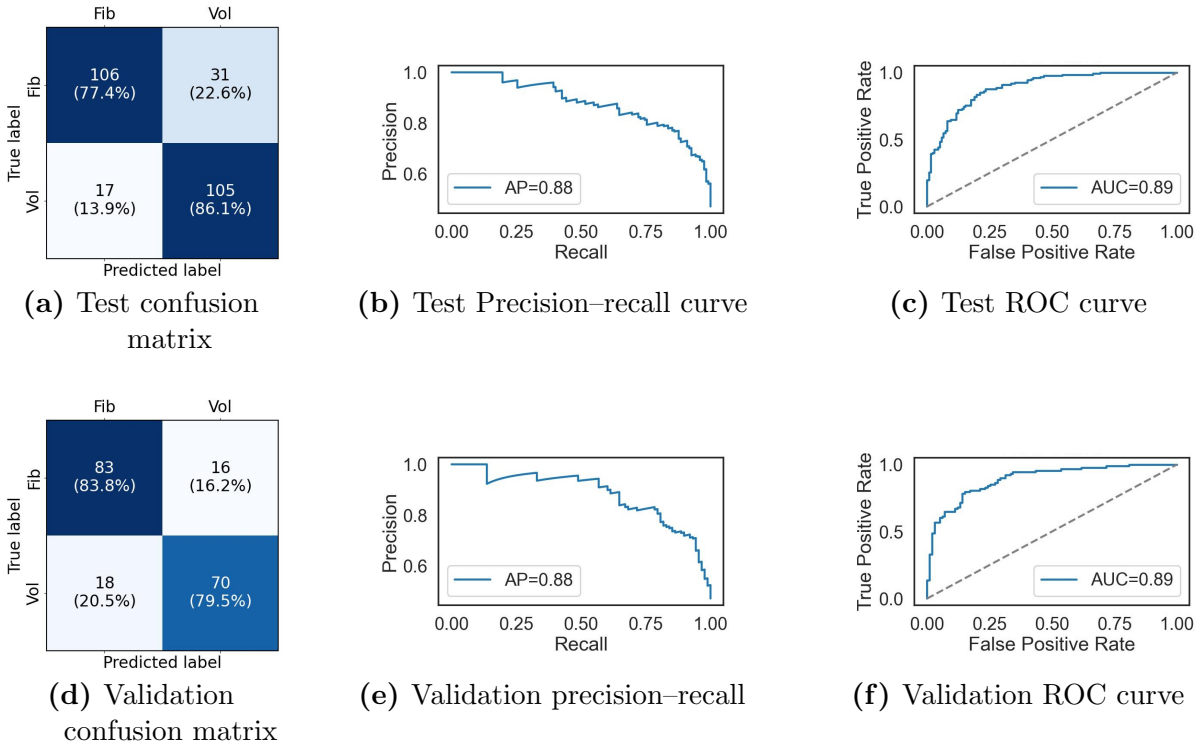
| Model Wavelet | Set | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Incept. Morse | Val. | 80.9 ± 0.076 | 80.9 ± 0.076 | 80.9 ± 0.076 | 80.0 ± 0.065 | 79.8 ± 0.071 | 0.88 ± 0.059 | 90.9 ± 0.034 | 0.59 ± 0.160 | 80.4 ± 0.085 | 0.76 ± 0.084 | 0.50 ± 0.088 |
| | Test | 81.1 ± 0.064 | 79.7 ± 0.068 | 81.9 ± 0.067 | 81.9 ± 0.067 | 79.5 ± 0.063 | 0.89 ± 0.037 | 89.7 ± 0.021 | 0.60 ± 0.108 | 79.7 ± 0.082 | 0.76 ± 0.082 | 0.49 ± 0.087 |
| Incept. Morlet | Val. | 74.7 ± 0.164 | 74.7 ± 0.164 | 74.7 ± 0.164 | 73.8 ± 0.164 | 73.1 ± 0.173 | 0.79 ± 0.198 | 82.9 ± 0.131 | 0.48 ± 0.321 | 74.7 ± 0.164 | 0.73 ± 0.116 | 0.52 ± 0.097 |
| | Test | 72.9 ± 0.111 | 73.0 ± 0.111 | 71.9 ± 0.103 | 71.9 ± 0.103 | 72.1 ± 0.107 | 0.82 ± 0.087 | 85.2 ± 0.051 | 0.45 ± 0.212 | 73.0 ± 0.111 | 0.72 ± 0.110 | 0.53 ± 0.090 |
| MaxViT Morse | Val. | 83.5 ± 0.071 | 83.5 ± 0.071 | 83.5 ± 0.071 | 83.2 ± 0.080 | 83.0 ± 0.077 | 0.89 ± 0.079 | 90.5 ± 0.049 | 0.68 ± 0.153 | 84.3 ± 0.070 | 0.80 ± 0.042 | 0.46 ± 0.048 |
| | Test | 86.2 ± 0.045 | 86.2 ± 0.049 | 85.5 ± 0.045 | 85.5 ± 0.045 | 85.5 ± 0.046 | 0.94 ± 0.035 | 94.9 ± 0.045 | 0.71 ± 0.092 | 86.2 ± 0.049 | 0.80 ± 0.015 | 0.46 ± 0.014 |
| MaxViT Morlet | Val. | 84.3 ± 0.070 | 84.3 ± 0.070 | 84.3 ± 0.070 | 85.4 ± 0.060 | 84.4 ± 0.066 | 0.95 ± 0.027 | 95.9 ± 0.029 | 0.70 ± 0.122 | 84.8 ± 0.067 | 0.80 ± 0.037 | 0.46 ± 0.048 |
| | Test | 82.2 ± 0.034 | 82.4 ± 0.026 | 81.5 ± 0.030 | 81.5 ± 0.030 | 81.4 ± 0.033 | 0.92 ± 0.021 | 92.8 ± 0.015 | 0.63 ± 0.062 | 82.4 ± 0.026 | 0.80 ± 0.025 | 0.46 ± 0.030 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1. = F1 Score; AP = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Mean Confidence; Ent. = Mean Entropy; Val. = Validation; Incept. = InceptionResNetV2. All metrics are reported as mean ± standard deviation across the five splits.

The observed variation in model performance across different data splits, demonstrated by the InceptionResNetV2 model on Morse data (highest performance on split 3: balanced accuracy 0.87, ROC-AUC 0.95, Cohen's kappa 0.75; lowest on split 2: balanced accuracy 0.73, ROC-AUC 0.84, Cohen's kappa 0.45), highlights inherent dataset variability. These performance fluctuations likely reflect differences in signal quality, noise levels, or patient-specific EMG patterns across the data subsets. This emphasizes the critical importance of employing cross-validation methods to achieve a robust and reliable assessment of model performance.

### Gradient-weighted Class Activation Mapping Analysis of Model Decisions

Inspection of the Grad-CAM overlays in Fig. V.37 confirms that the network relies on different time-frequency patterns when separating fibrillation potentials from voluntary contractions. The model's attention consistently highlights specific morphological characteristics in scalograms that appear to be the primary drivers for its classification decisions.




**(a)** Test confusion matrix

**(b)** Test Precision–recall curve

**(c)** Test ROC curve

**(d)** Validation confusion matrix

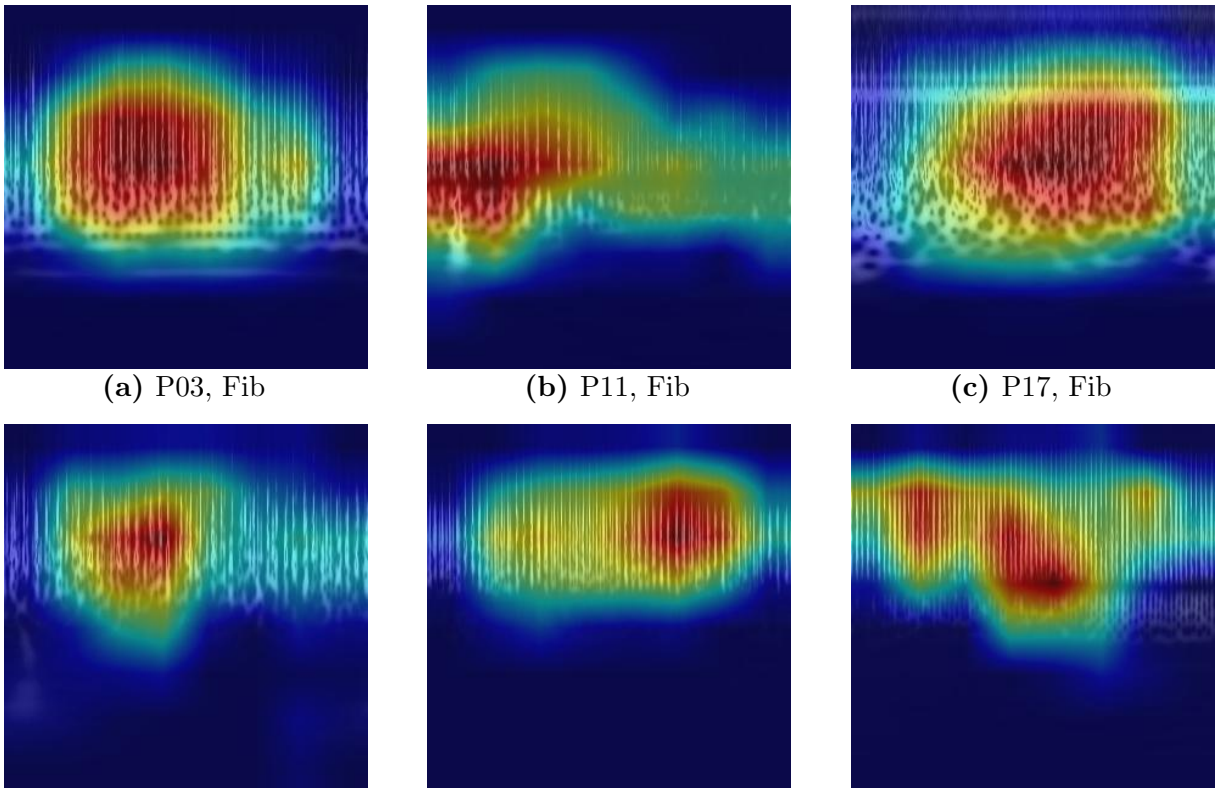
**(e)** Validation precision–recall

**(f)** Validation ROC curve

**Figure V.36:** Aggregated classification performance of the InceptionResNetV2 model on the Morse-wavelet-transformed dataset, averaged across all splits.

For instances correctly classified as fibrillation potentials (Fig. V.37a, Fig. V.37b, Fig. V.37c), the salient regions identified by Grad-CAM are predominantly characterized by their spectrally broad nature. The vertical extent of these hotspots on the scalogram indicates activation across a wide range of frequencies, extending from lower frequencies well into the upper portions of the spectrum. This shows that the model associates Fib class with events that possess significant energy across diverse frequency components.

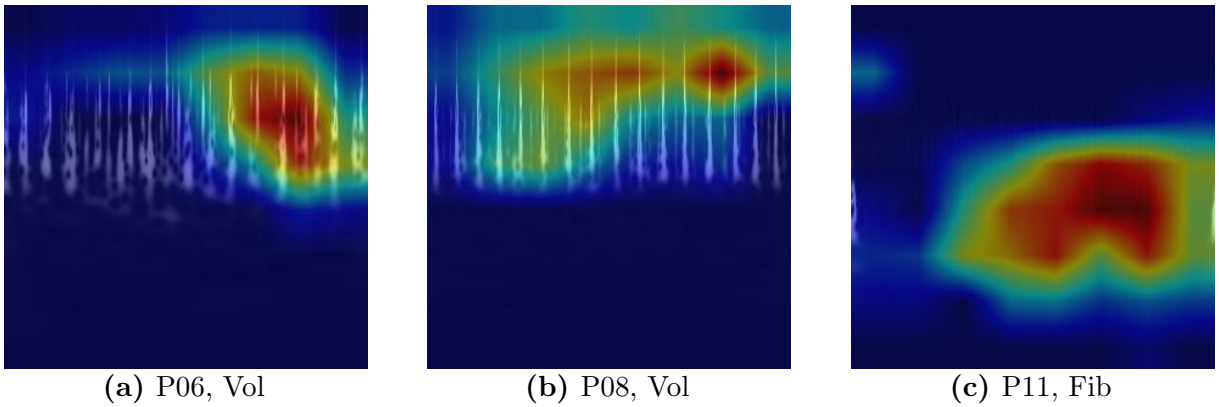
In contrast, for recordings correctly classified as voluntary contraction potentials (Fig. V.37d, Fig. V.37e, Fig. V.37f), the Grad-CAMs highlight regions with a different morphology. The salient features for Vol class are characterized by their temporally sustained presence and confinement to a relatively narrow band of higher frequencies. These activations typically appear as elongated horizontal "ribbons" that span a substantial part of the analysis window's duration. Vertically, these ribbons are concentrated in the upper portion of the scalogram, indicating that the model primarily



**(a)** P03, Fib **(b)** P11, Fib **(c)** P17, Fib

**(d)** P06, Vol **(e)** P08, Vol **(f)** P10, Vol

**Figure V.37:** Grad-CAM overlays for six correctly predicted recordings (**top**: fibrillation potentials, **bottom**: voluntary contraction potentials). Hot areas for fibrillation shrink into brief, broadband bursts, whereas voluntary contraction saliency forms a sustained, narrow-band ridge.



**(a)** P06, Vol **(b)** P08, Vol **(c)** P11, Fib

**Figure V.38:** Grad-CAM overlays for three recordings that the CNN misclassified. In the two **Vol** cases (left, middle), the late broadband burst draws saliency away from the faint stationary ridge and triggers a **Fib** decision; in the **Fib** case (right), the recurrent low-band activity produces a sustained, narrow-band saliency ridge, steering the network toward the **Vol** class.

attends to persistent energy in the upper frequency ranges for identifying voluntary activity.

The analysis of misclassified samples further clarifies the model's decision logic and its failure modes: When voluntary signals were misclassified as fibrillation (Fig. V.38a, Fig. V.38b), the Grad-CAMs reveal that the model's attention was captured by features within the window that, while atypical for Vol class, resembled the Fib class patterns. In Fig. V.38a, the dominant salient region is a temporally localized, broadband hotspot similar to typical Fib CAMs. In Fig. V.38b, the CAM focuses on a feature that is more broadband and less temporally sustained than classic Vol patterns, again aligning its morphology more with the Fib class. The model, therefore, bases its incorrect decision on these isolated, ambiguous Fib-like patterns.

When a fibrillation potential is labeled as voluntary action potential (Fig. V.38c), the salient area is broad in time, resembling the sustained activity of a voluntary contraction, but its spectral span is wider and more irregular than is usual for true Vol maps. The pattern thus blends the temporal persistence associated with voluntary firing and the spectral heterogeneity expected in fibrillation. The classifier appears to weight the temporal cue more strongly, leading to a voluntary verdict despite the presence of higher-frequency content.

Across the errors the Grad-CAMs do not exactly replicate a canonical template of the opposite class; instead they present compound saliency patches that share frequency characteristics with one class and temporal characteristics with the other. These ambiguous signatures blur the decision boundary learned during training and prompt the model to favor whichever cue – temporal extent or spectral width, carries slightly more integrated saliency. So misclassifications occur when a single window contains overlapping features from both class morphologies.

**Ensemble Performance**

Pooling the predictions of all eight base learners yields a well-rounded classifier whose behavior is summarized in Tab. V.11 and visualized in Fig. V.39. The ensemble attains an average accuracy of 86.9 % with a virtually identical balanced accuracy, indicating that sensitivity to fibrillations (87.1 %) and specificity to voluntary potentials (86.1 %) are tightly matched. Discrimination is strong: the ROC-AUC of 0.93 and the average precision of 94.2 % confirm that the ranking of positive versus negative instances is consistently reliable across splits. Agreement with ground truth is substantial ($\kappa = 0.73$), and the negative-log-likelihood and Brier score point to well-shaped probability outputs.

The aggregated confusion matrix (Fig. V.39a) corroborates these findings, showing that only 12.4 % of fibrillation events and 13.9 % of normal discharges are misclassified. Precision remains above 0.9 until recall approaches 0.8 (Fig. V.39b), while the ROC

curve maintains a high true-positive rate even at low false-positive rates (Fig. V.39c). Calibration analysis (Fig. V.39d) reveals a slight over-confidence for probabilities between 0.5 and 0.8; nevertheless, the expected-calibration error is modest (ECE $= 0.10 \pm 0.06$). Selective-prediction performance rises sharply with confidence (Fig. V.39e), with accuracy surpassing 95 % at confidence thresholds above approximately 0.75, and the entropy distribution (Fig. V.39f) shows that the majority of samples cluster in a low-to-moderate uncertainty regime.

**Table V.11:** Aggregated Metrics for Ensemble Classification of Fibrillation Potentials across Splits

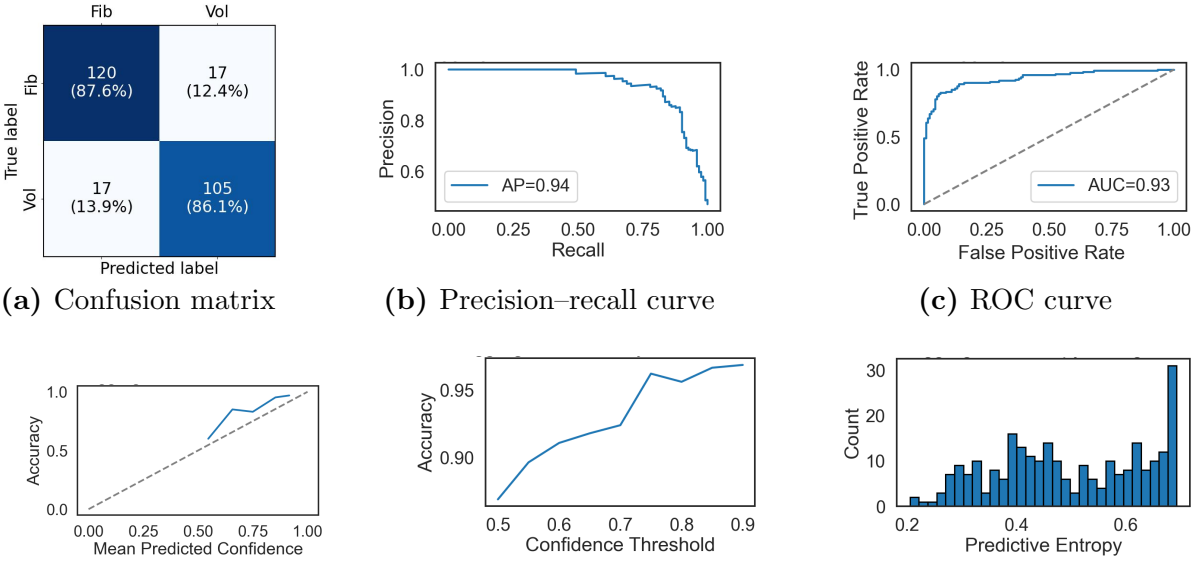
| Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. | NLL | Brier Score | ECE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 86.9 ± 0.082 | 86.8 ± 0.087 | 87.1 ± 0.087 | 86.8 ± 0.082 | 86.8 ± 0.084 | 0.93 ± 0.044 | 94.2 ± 0.037 | 0.73 ± 0.168 | 86.8 ± 0.087 | 0.771± 0.032 | 0.497± 0.038 | 0.369± 0.061 | 0.109± 0.027 | 0.098± 0.056 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1. = F1 Score; AP. = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Mean Confidence; Ent. = Mean Entropy. All metrics are reported as mean ± standard deviation across the five splits.

Despite the narrow overall standard deviations, some of individual splits display pronounced heterogeneity (Tab. V.12). The best ensemble performance was registered on split 4, which is classified almost flawlessly. The model correctly labels every fibrillation potential and all but one voluntary unit (Fig. V.40a), giving a balanced accuracy of 97.4 %, sensitivity of 100 % and specificity of 94.7 %. Fig. V.40b and Fig. V.40c show that discrimination saturates (ROC-AUC $= 1.00$, AP $= 1.00$), while reliability analysis (Fig. V.40d) indicates that the model is mildly under-confident in the 0.55-0.75 probability range: observed accuracy exceeds predicted confidence, reflected in a mean confidence of 0.72, entropy of 0.55 and an ECE of 0.25. The selective-prediction curve (Fig. V.40e) reaches almost 100 % accuracy once predictions with confidence below 0.75 are discarded. These results indicate that the ensemble separates the two classes with near-perfect effectiveness.

Split 2 presents the most challenging data subset for the ensemble. Out of 56 samples it misclassified six fibrillation events and eight voluntary units (Fig. V.41a), yielding a balanced accuracy of 74.8 %. Sensitivity to Fib is 73.9 % and specificity to Vol is 75.8 %; nevertheless, the ROC-AUC stays reasonably high at 0.88 and the AP reaches 0.92, indicating that the ranking function is still effective even though the default decision threshold is sub-optimal. Confidence diagnostics corroborate this view: the reliability curve (Fig. V.41d) shows slight under-confidence for probabilities below 0.8, while the selective-prediction curve (Fig. V.41e) surpasses 90 % accuracy once predictions with confidence below 0.75 are rejected and attains near-perfect accuracy at the 0.9 threshold. The broad entropy distribution (Fig. V.41f) confirms that the ensemble recognizes the

ambiguity of these borderline waveforms, so performance could be markedly improved by deferring low-confidence cases to re-evaluation.




**(a)** Confusion matrix **(b)** Precision–recall curve **(c)** ROC curve

**(d)** Reliability curve **(e)** Selective-prediction curve **(f)** Distribution of entropy

**Figure V.39:** Aggregated results for the ensembled classifier distinguishing fibrillation vs. voluntary action potentials across all test splits. **Top row**: confusion and discrimination metrics. **Bottom row**: calibration and uncertainty analyses. These complement the numerical metrics in Tab. V.11.

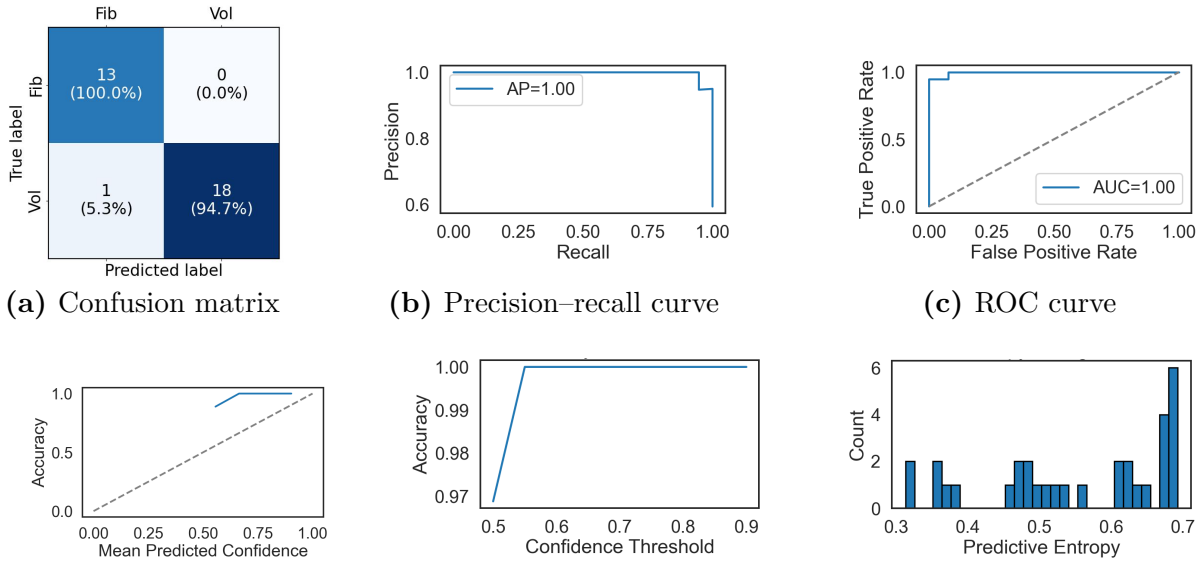
**Table V.12:** Metrics for Ensemble Predictions for Splits 2 and 4

| Split # | Acc. | Rec. | Spec. | Prec. | F1 | ROC AUC | AP | $\kappa$ | Bal. Acc. | Conf. | Ent. | NLL | Brier Score | ECE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 75.0 | 73.9 | 75.8 | 74.3 | 74.5 | 0.88 | 92.5 | 0.49 | 74.8 | 0.80 | 0.47 | 0.48 | 0.16 | 0.07 |
| 4 | 96.9 | 100 | 94.7 | 96.4 | 96.8 | 1.00 | 99.7 | 0.94 | 97.4 | 0.72 | 0.55 | 0.35 | 0.10 | 0.25 |

**Note:** Acc. = Accuracy; Rec. = Recall; Spec. = Specificity; Prec. = Precision; F1. = F1 Score; AP. = Average Precision; Bal. Acc. = Balanced Accuracy; Conf. = Mean Confidence; Ent. = Mean Entropy. All metrics are reported as mean ± standard deviation across the five splits.

We also evaluated an alternative ensemble configuration by excluding the weakest-performing model – InceptionResNetV2 trained on Morlet-wavelet data and retaining only the three best-performing models. This smaller ensemble showed improved classification accuracy, reaching 87.6 %, albeit with a slight degradation in calibration metrics reflecting increased model uncertainty. This trade-off highlights the common tension between diversity and calibration quality in ensemble construction.

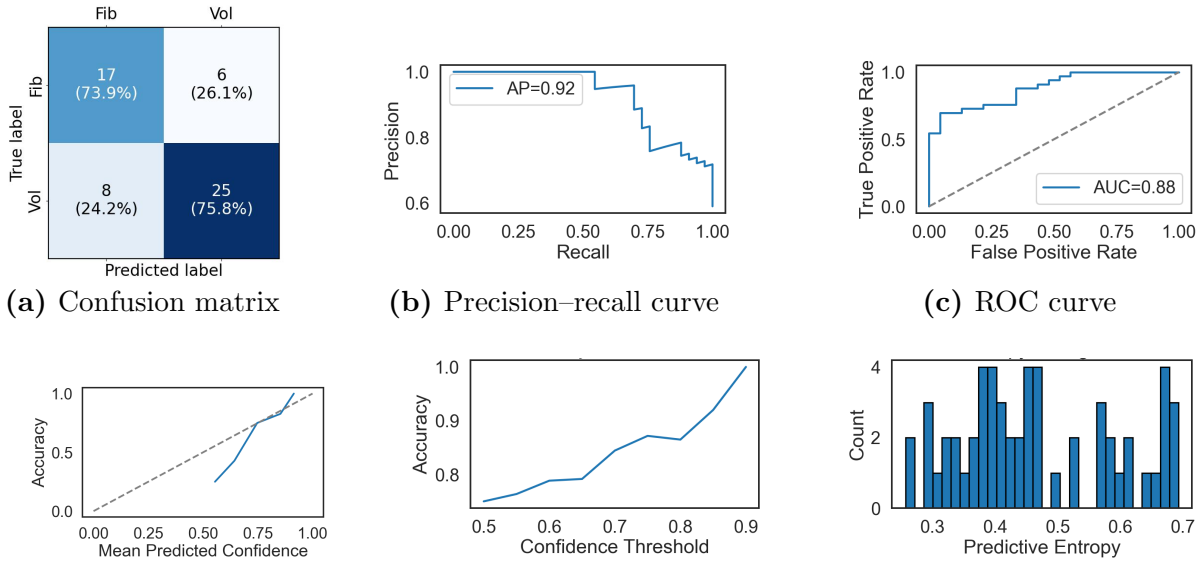
Overall, ensembling significantly reduces split-to-split variability seen in individual models, enhancing discrimination metrics to levels potentially suitable for clinical application.




**(a)** Confusion matrix **(b)** Precision–recall curve **(c)** ROC curve

**(d)** Reliability curve **(e)** Selective-prediction curve **(f)** Distribution of entropy

**Figure V.40:** Performance and uncertainty diagnostics for the split-4 (best) model distinguishing fibrillation vs. voluntary action potentials. **Top row**: confusion and discrimination. **Bottom row**: calibration an uncertainty. These complement the numerical metrics in Tab. V.12.




**(a)** Confusion matrix **(b)** Precision–recall curve **(c)** ROC curve

**(d)** Reliability curve **(e)** Selective-prediction curve **(f)** Distribution of entropy

**Figure V.41:** Performance and uncertainty diagnostics for the split-2 (worst) model distinguishing fibrillation vs. voluntary action potentials. **Top row**: confusion and discrimination. **Bottom row**: calibration and uncertainty. These complement the numerical metrics in Tab. V.12.

However, residual calibration issues, reflected by slight overconfidence, as well as substantially lower performance on the most challenging splits, suggest the need for improved strategies to handle ambiguous or borderline cases. Applying post-ensemble calibration methods such as temperature scaling or isotonic regression, along with uncertainty-aware loss functions could mitigate remaining biases and further enhance generalization across diverse patient datasets.

## V.4 Obstructive Sleep Apnea Dataset: Classification

In this section, we test our hypothesis that the shape of patient airflow is predictive of a successful response to MAS treatment in OSA patients. For this dataset, minimizing false negatives is of utmost clinical importance, as overlooking a potential positive case – responders, could result in a missed treatment opportunity. Consequently, recall will be prioritized as the primary evaluation metric. We begin by applying an image-based transfer learning pipeline that adapts a pre-trained ImageNet model to extract deep features from time–frequency representations of the airflow signals – a method that has proven effective in our previous application on skeletal muscle disorder datasets.

Next, we complement this deep learning approach with a statistical texture analysis of the scalograms using GLCM features. This step quantitatively assesses inherent textural patterns within the scalograms, providing an independent validation regarding the discriminative potential of airflow signals in differentiating between classes.

Finally, we employ classical machine learning methods on time-domain and frequency-domain features extracted from the raw signals (see Subsection IV.6.1). By comparing the performance of various classifiers on these features, we evaluate the effect of sampling frequency on the models' ability to discriminate between the two patient groups.

### V.4.1 Application of Scalogram-Based Transfer Learning to Polysomnography Data Classification

Our initial effort focuses on extending the transfer learning pipeline, previously validated on skeletal muscle disorder datasets, to the classification of OSA patients. To ensure robust evaluation, we divide the dataset into five grouped cross-validation splits, stratified by responder status, such that all images from an individual patient appeared exclusively in either the training, validation, or test set. An InceptionResNetV2 model is trained on individual scalogram images obtained from signal segments as described in Subsection IV.6.2.

The performance metrics summarized in the Row 1 of Tab. V.13 reveal significant challenges in the model's ability to generalize. Analysis of the training and validation curves (Fig. V.44a and Fig. V.44b) indicates that, although training loss is consistently low, it shows minimal improvement across epochs. Conversely, validation loss exhibits

significant instability, characterized by fluctuations that suggest ineffective model learning and potential overfitting or inability to capture discriminative features. Indeed, the validation metrics reflect these concerns, with an average accuracy below 55 %, specificity as low as approximately 36 % in the baseline approach, only marginally above-chance recall, and notably a negative Cohen's Kappa for the baseline, indicating performance worse than random predictions relative to class distribution.

Given the limited number of patients, improving the network's generalization is essential. One practical method is to apply hierarchical or adaptive pooling, summarized in Algorithm 3. Adaptive pooling aggregates multiple epoch-level scalogram images from each patient into a single global scalogram. This aggregation, achieved via statistical operations such as average, max, or percentile pooling, captures the overall spectral content and temporal variability of the data, significantly reducing its dimensionality.

The primary advantage of this approach is its simplicity. With one aggregated image per patient, we can leverage pre-trained convolutional neural networks directly, without extensive modifications to account for temporal dynamics. This reduction in dimensionality not only simplifies the classification pipeline but also reduces the number of trainable parameters, thereby mitigating the risk of overfitting, a particularly important consideration given the small sample size.

However, there is an inherent trade-off: aggregating scalograms across epochs results in the loss of fine temporal resolution. This smoothing effect may obscure transient or nonstationary patterns that might be crucial for differentiating between classes. Thus, the choice of pooling statistic is critical. For instance, average pooling can provide a robust, noise-reducing summary, whereas max or percentile pooling might better preserve extreme values indicative of significant events

For implementation of this method, we first employ a pre-trained InceptionResNetV2 backbone (with its classifier head removed) to extract deep features from every scalogram image. We then apply an adaptive pooling step, implemented as an average across all segment-level features for each patient, to obtain a global representation. Finally, a simple linear classifier is trained on these pooled features within a cross-validation framework.

The metrics in Tab. V.13 show that the adaptive pooling approach yields an accuracy of about 42 %, a recall near 17 %, and an F1 score of roughly 0.16, indicating weaker overall performance than the baseline model in most respects. Notably, the specificity (64 %) is higher than in the baseline, implying a stronger bias toward the negative class. This shift in class prediction leads to a lower recall and a negative Cohen's Kappa, suggesting that the pipeline struggles to correctly identify responders.

**Algorithm 3** Patient-Level Classification Pipeline with Adaptive Pooling

**Require:** Patient dataset $\mathcal{D} = \{(p_i, y_i)\}_{i=1}^{N}$, split configuration, pretrained feature extractor $\mathcal{F}$, pooling method $\mathcal{P}$, where each $p_i$ is a folder containing scalogram images and $y_i \in \{0, 1\}$ is the corresponding label.

Load patient splits from the provided JSON configuration.
Initialize $\mathcal{F}$, freeze its parameters, and set it to `eval` mode.
Determine feature dimension $d$ using a dummy input.
**for** each fold in the cross-validation splits **do**
  Split $\mathcal{D}$ into training set $\mathcal{D}_{train}$ and validation set $\mathcal{D}_{val}$.
  **for** each patient $(p, y) \in \mathcal{D}_{train} \cup \mathcal{D}_{val}$ **do**
    Retrieve sorted images $I_p = \{I_{p,1}, \ldots, I_{p,n}\}$ from folder $p$.
    Apply transformation $\mathcal{T}$ to each image in $I_p$.
    **for** $j = 1$ **to** $n$ **do**
      Compute feature: $f_{p,j} \leftarrow \mathcal{F}(I_{p,j})$.
    **end for**
    Aggregate features: $f_p^{agg} \leftarrow \mathcal{P}(\{f_{p,j}\}_{j=1}^{n})$.
  **end for**
  Construct Data Loaders for training and validation.
  Initialize classifier $\mathcal{C}$ (linear layer: $\mathbb{R}^d \to \mathbb{R}^2$).
  Define loss $\mathcal{L}$ (Cross-Entropy with label smoothing) and optimizer (AdamW).
  **for** each epoch **do**
    **Train:** For each batch $(f_p^{agg}, y)$, compute $\hat{y} \leftarrow \mathcal{C}(f_p^{agg})$ and update parameters by minimizing $\mathcal{L}(\hat{y}, y)$.
    **Validate:** Evaluate performance and record metrics.
    Save checkpoint if validation loss improves.
  **end for**
  Record fold metrics.
**end for**
Compute average metrics and aggregate confusion matrix across folds.
Save final metrics and plots.

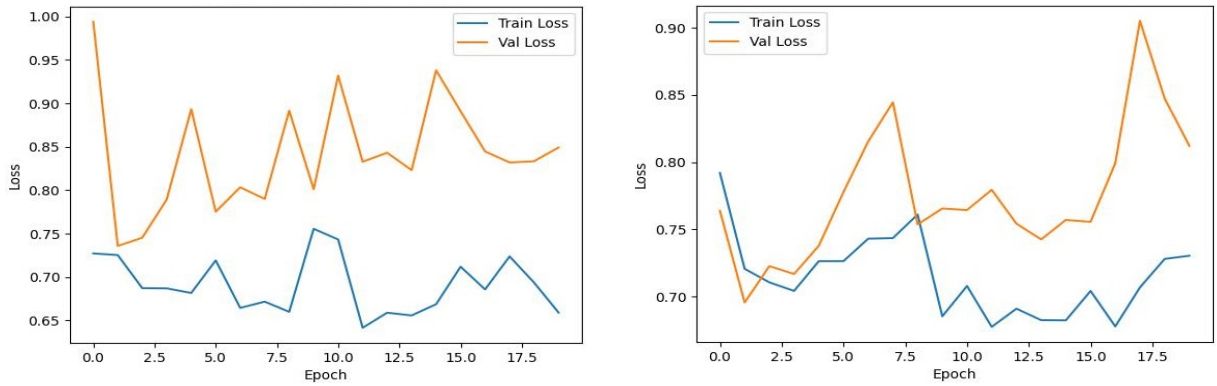
**Table V.13:** Classification Metrics for Baseline, Adaptive Pooling, Feature Fusion, and Fine-Tuning, Simple CNN with Feature Fusion Pipelines on Validation Set

| Model | Acc. | Recall | $F_\beta$ Score | Prec. | $\kappa$ | Bal. Acc. | Spec. | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| Baseline | 43.6 | 55.4 | 45.1 | 39.3 | −0.088 | 45.7 | 36.1 | 0.469 |
| Adaptive Pooling | 41.8 | 17.1 | 15.6 | 14.7 | −0.186 | 40.8 | 64.4 | 0.370 |
| Feature Fusion | 54.2 | 45.7 | 44.6 | 55.6 | 0.079 | 53.7 | 61.7 | 0.502 |
| Fine-Tuning | 48.2 | 8.6 | 12.3 | 32.9 | −0.090 | 45.4 | 82.2 | 0.538 |
| Simple CNN + Feature Fusion | 54.3 | 0.0 | 0.0 | 0.0 | 0.0 | 50.0 | 100.0 | 0.467 |

**Note:** Acc. = Accuracy; Prec. = Precision; Spec. = Specificity; Bal. Acc. = Balanced Accuracy.

Fig. V.42 presents representative training and validation loss curves for two cross-validation folds under this adaptive pooling setup. Although the training loss decreases, the validation

loss remains volatile and does not converge to a distinctly lower level. This pattern suggests that the global feature representation may not sufficiently capture patient-specific temporal variations, and the small dataset size likely amplifies these generalization challenges.



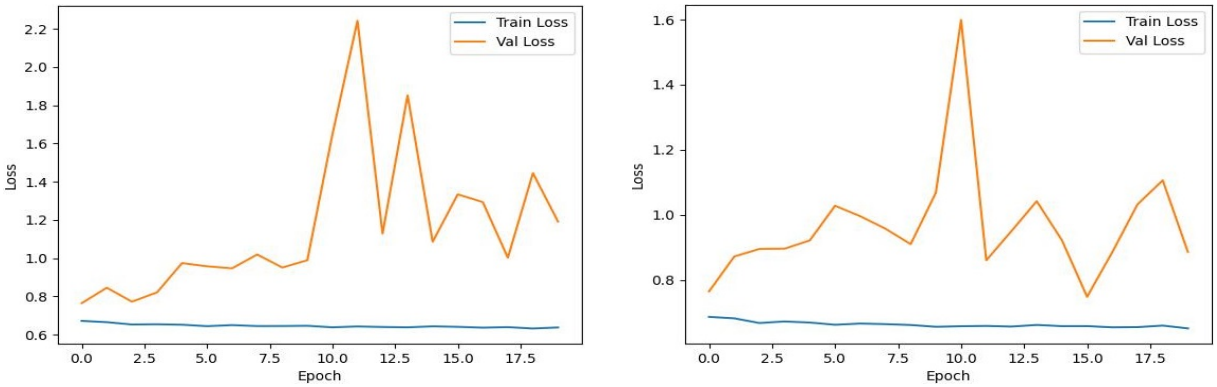

**(a)** Training/Validation Loss – Fold 1 **(b)** Training/Validation Loss – Fold 2

**Figure V.42:** Representative training and validation loss curves for two cross-validation folds under the adaptive pooling pipeline. The inconsistent validation loss indicates that the aggregated features may not reliably discriminate between responders and non-responders.

CNN-RNN fusion models combine the spatial feature extraction capability of CNN with the temporal sequence modeling power of recurrent neural networks (RNN), such as long short-term memory networks. In a conventional CNN-RNN architecture, the CNN first processes raw signals, typically in the form of time series or time-frequency representations, to identify local patterns and hierarchical features, including waveform shapes and spectral content. These extracted features are then organized into sequences and passed to an RNN, which learns the temporal dependencies among them to model long-term dynamics and contextual evolution. Finally, fully connected layers or a softmax classifier transform the RNN's output into class predictions. This end-to-end design allows simultaneous optimization of both convolutional filters and recurrent units, thereby enhancing classification performance.

For our application, we adapt this approach, described in the Algorithm 4, to an image classification task by substituting the conventional CNN with a network pre-trained on ImageNet and using scalogram images derived from time-series data. In our revised pipeline, each patient's recording is processed as a sequence of scalogram segments. Specifically, each patient's folder is loaded and its images are standardized to 107 segments by truncating excess images or padding with zeros when necessary. Each scalogram segment is then passed through a pre-trained InceptionResNetV2 backbone to extract high-level features. These feature vectors are subsequently arranged into a sequential order and fed into an LSTM to fuse the temporal information. The final hidden state of the LSTM is used by a fully connected layer to predict the patient's class. The training loop employs cross-entropy

loss with label smoothing, while standard evaluation metrics, model checkpoints, and training histories are recorded across cross-validation folds. This sequential fusion approach is designed to capture the temporal dynamics across recording segments more effectively than simple pooling methods.




**(a)** Training/Validation Loss – Fold 1 **(b)** Training/Validation Loss – Fold 2

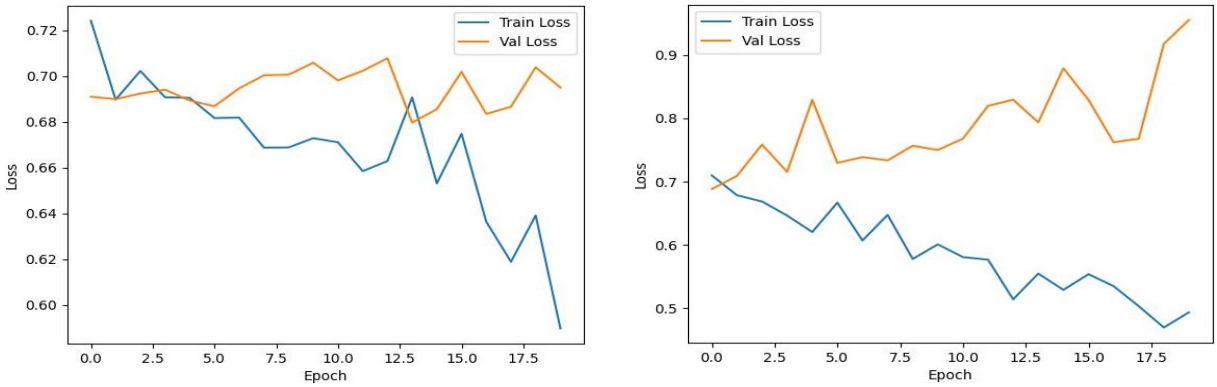
**Figure V.43:** Training and validation loss curves for two representative cross-validation folds (Fold 1 and Fold 2) for baseline model. The plots illustrate how the model's training and validation performance evolves over the epochs, indicating potential overfitting behavior.

Our evaluation of the sequential feature fusion pipeline reveals only a marginal improvement over chance-level classification: refer to the Row 3 of Tab. V.13 for the performance metrics. The model attains an overall accuracy of approximately 54 % and a recall of around 46 %, Although it now makes some positive predictions, unlike earlier versions that predominantly predicted the negative class, its discriminative power remains limited. The balanced accuracy barely exceeds 50 % and Cohen's Kappa is near zero, indicating that, once class imbalance is considered, the predictions are nearly equivalent to random guessing.

These results indicate that, although the model does learn some features, the data do not exhibit strongly separable patterns between responders and non-responders. The near-chance AUC of 0.50 suggests that the predicted probabilities do not effectively rank patients by class. This limited performance may stem from factors such as inherent noisiness, subtle differences in the scalogram data, the small sample size, and the high variability in signals. Additionally, the frozen pre-trained CNN backbone may not fully capture domain-specific characteristics, which highlights the need for additional data or partial fine-tuning of the backbone.

Fig. V.44a and Fig. V.44b present the training and validation loss curves for two representative cross-validation folds under the feature fusion pipeline. Both folds exhibit considerable fluctuations in loss across epochs, indicating that the model does not settle into a consistently lower-loss regime. The training loss occasionally decreases but frequently

rebounds, while the validation loss similarly varies without a clear downward trend. This volatility may reflect difficulties in learning stable features from a limited, noisy dataset or in adapting the pre-trained backbone to the nuances of OSA-related scalograms. Moreover, the final training and validation losses remain relatively close, suggesting that the model is neither severely overfitting nor achieving robust generalization. Overall, these observations imply that the current architecture and hyperparameters do not fully capture the differences between responders and non-responders, emphasizing the need for additional regularization, more data, or domain-specific fine-tuning to attain stronger and more stable performance.




**(a)** Training/Validation Loss – Fold 1 **(b)** Training/Validation Loss – Fold 2

**Figure V.44:** Training and validation loss curves for two representative cross-validation folds (Fold 1 and Fold 2) under the feature fusion pipeline. The plots illustrate how the model's training and validation performance evolves over the epochs, highlighting challenges with convergence.

We next unfreeze select layers of the pre-trained InceptionResNetV2 model to fine-tune it on our OSA dataset. The motivation is to adapt the network's higher-level filters to domain-specific patterns while still leveraging the lower-level feature representations learned on ImageNet. In practice, we unfreeze the later blocks while keeping earlier layers frozen, aiming to balance adaptation with the risk of overfitting.

Despite these adjustments, the results summarized in the Row 4 of Tab. V.13 reveal only marginal improvements in discriminating responder from non-responder classes. Specifically, the accuracy remains below 50 %, with recall at approximately 8 % and specificity exceeding 80 %. This imbalance leads to a low F1 score (about 0.12) for the positive class, reflecting that the model rarely identifies true positives. Although the ROC-AUC (around 0.54) suggests some weak separation in predicted probabilities, the low recall indicates that the network effectively misses most responder cases. The negative Cohen's Kappa underscores that it performs near or below random chance once class distribution is taken into account.

Comparing these results with earlier attempts, there is little improvement in correctly

classifying the positive class: the model still leans heavily on predicting negatives. These findings echo previous attempts in that there is little improvement in capturing the positive class. The strong negative-class bias suggests that the limited size and potentially subtle nature of the underlying scalogram features do not provide a robust signal for the network to exploit.

For our final test, we train a custom convolutional backbone from scratch to extract segment-level embeddings, which are subsequently fused by an LSTM for patient-level classification. The pipeline processes each patient's 107 scalogram segments individually through the CNN, then arranges the resulting feature vectors in a temporal sequence for the LSTM.

Despite achieving an overall accuracy of about 54 %, the model's recall is effectively zero, indicating that it consistently fails to identify the responder class. This is reflected in the perfect specificity and a balanced accuracy of 50 %. The model's decision boundary collapses to predicting all patients as non-responders, which yields no true positives. This behavior suggests that the learned features do not sufficiently differentiate responder signals, possibly due to limited data, subtle class-specific patterns, or insufficient architectural complexity in the CNN for capturing the relevant spectral signatures. The resulting low ROC-AUC (about 0.47) further confirms that the predicted probabilities offer little discriminatory power. Overall, the model predicts exclusively the negative class and fails to generalize to the minority positive class.

One plausible explanation for the pipeline's under-performance is that the polysomnographic data, and the corresponding scalograms derived from them, do not exhibit consistent or discriminative visual patterns for the ImageNet-trained network to extract meaningful features. These signals are highly variable across patients, and the conversion into scalogram images may further dilute critical temporal details. In addition, the low sampling frequency of the data, as discussed in Subsection IV.6.1, can further obscure discriminative characteristics. As a consequence, the network struggles to detect the subtle distinctions between responders and non-responders unless these differences are stark and consistent.

Another contributing factor is the quality and balance of the dataset. The relatively small sample size, combined with class imbalance and the presence of noise or artifacts in the images, may cause the model to overfit on trivial cues present in the training set while failing to generalize to new examples. In particular, if data preprocessing or splitting introduces overlap or unintended biases between the training and validation sets, the model may learn spurious features that do not persist in unseen data. This is reflected in the pronounced training-validation discrepancy and the overall low recall, indicating that the network effectively defaults to predicting the majority negative class.

If the underlying signals truly lack visually distinctive patterns or if the critical information is lost during the transformation into scalograms a purely image-based deep learning approach may be fundamentally ill-suited for this task. Taken together, these observations underscore the need to address limitations related to data variability and insufficient discriminative information. In the upcoming subsection, we will further investigate whether the scalograms indeed do not contain class-discriminative features by applying texture analysis methods.

**Algorithm 4** Patient-Level Classification Pipeline with Sequential Feature Fusion

**Require:** Patient dataset $\mathcal{D} = \{(p_i, y_i)\}_{i=1}^{N}$, pretrained feature extractor $\mathcal{F}$, sequence model (LSTM), and fully connected layer classifier, where each $p_i$ is a folder containing scalogram images and $y_i \in \{0, 1\}$ is the corresponding label.

Set `MAX_SEGMENTS` $\leftarrow 107$.
**for** each patient $(p, y) \in \mathcal{D}$ **do**
  Retrieve sorted images $I_p = \{I_{p,1}, \ldots, I_{p,n}\}$ from folder $p$.
  **if** $|I_p| >$ `MAX_SEGMENTS` **then**
    Truncate $I_p$ to the first `MAX_SEGMENTS` images.
  **else if** $|I_p| <$ `MAX_SEGMENTS` **then**
    Pad $I_p$ with zero images to reach `MAX_SEGMENTS`.
  **end if**
  Apply transformation $\mathcal{T}$ to each image in $I_p$.
  Stack images to form tensor $X_p \in \mathbb{R}^{\texttt{MAX\_SEGMENTS} \times 3 \times H \times W}$.
**end for**
Construct a DataLoader yielding batches of patient tensors.
Initialize loss function $\mathcal{L}$ (Cross-Entropy) and optimizer.
**for** each training epoch **do**
  **for** each batch $\{(X_p, y_p)\}_{p=1}^{B}$ **do**
    **for** each patient $p$ in the batch **do**
      **for** $j = 1$ **to** `MAX_SEGMENTS` **do**
        Compute feature: $f_{p,j} \leftarrow \mathcal{F}(X_p[j])$.
      **end for**
      Stack features to form sequence $F_p \in \mathbb{R}^{\texttt{MAX\_SEGMENTS} \times D}$.
    **end for**
    Stack patient sequences to form tensor $F \in \mathbb{R}^{B \times \texttt{MAX\_SEGMENTS} \times D}$.
    Compute hidden representations: $H \leftarrow \mathrm{LSTM}(F)$.
    Extract final hidden state $h_{\text{final}}$ from $H$.
    Compute predictions: $\hat{y} \leftarrow \mathrm{FC}(h_{\text{final}})$.
    Compute loss: $\mathcal{L}(\hat{y}, y)$.
    Update model parameters via backpropagation.
  **end for**
  Validate the model on the validation set.
  Save checkpoint if validation performance improves.
**end for**
Save final metrics and plots.

### V.4.2 Statistical Analysis of Scalograms Based on Gray-Level Co-occurrence Matrix Features

Our initial attempts to classify scalogram images using transfer learning resulted in performance near random chance. We shifted our focus toward analyzing the inherent texture properties of these images to determine whether they contain discriminative information for differentiating treatment responders from non-responders. We extracted texture metrics using the GLCM method (see Section IV.4) and examined their relationship with clinical outcomes.

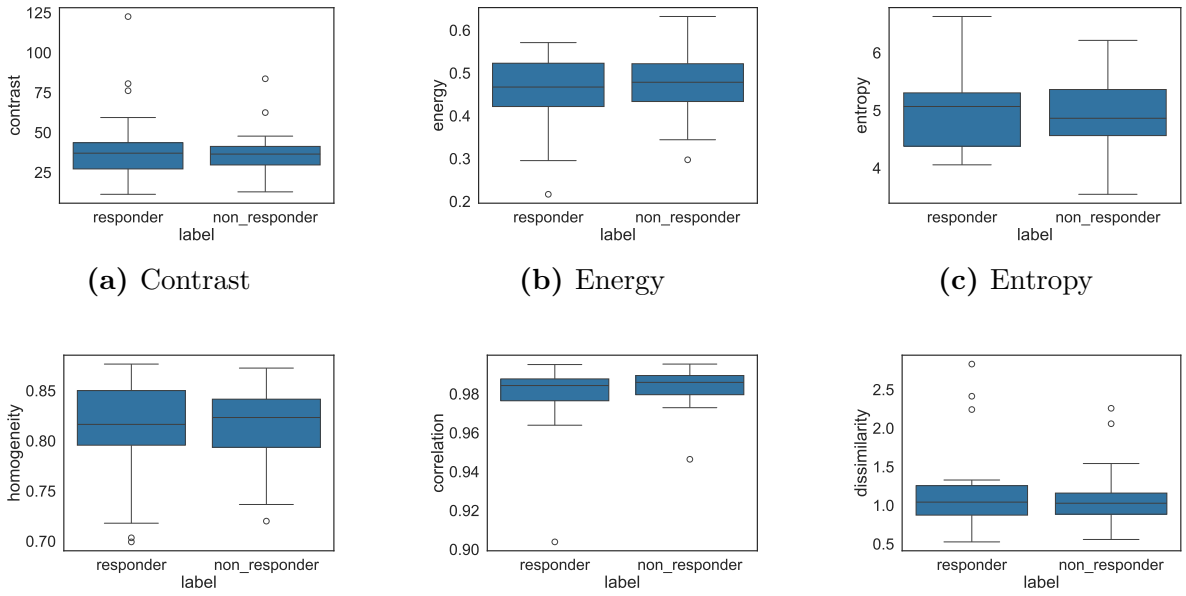
The analysis pipeline, implemented in Python (Python Core Team, 2019), converts each scalogram image to grayscale and computes its GLCM to extract the following texture features: contrast, energy, entropy, homogeneity, correlation, and dissimilarity. For each subject, multiple images are processed and the features are averaged to obtain subject-level descriptors, with responder status inferred from folder names. Statistical tests, such as Welch's t-test, Mann-Whitney U test, and permutation tests were applied to assess differences between groups, and effect sizes were quantified using Cohen's d. In parallel, boxplots were generated to visually compare the distributions of these features.

In the OSAMAS dataset (Fig. V.45), the boxplots for contrast, energy, and entropy as well as homogeneity, correlation, and dissimilarity reveal largely overlapping distributions between responders and non-responders. For instance, the contrast boxplot shows nearly identical medians and substantial overlap in the interquartile ranges of both groups. Quantitatively, contrast yields a t-statistic of 0.98 ($p = 0.33$), a Mann-Whitney U test p-value of 0.78, and a permutation test p-value of 0.31, with Cohen's d of 0.33, indicating only a small-to-moderate effect size. Similarly, energy exhibits a t-statistic of $-0.85$ with p-values around 0.40 (t-test), 0.70 (Mann–Whitney U), and 0.38 (permutation), and a Cohen's d of $-0.28$. The remaining features also produce p-values well above 0.05 and effect sizes with absolute Cohen's d values below 0.33, reinforcing that the texture measures do not significantly differentiate the two classes.

The results for CRC dataset are summarized in Fig. V.46. Here, the contrast boxplot suggests that responders might have a slightly higher median and broader range compared to non-responders, although the overall distributions still largely overlap. Specifically, contrast shows a t-statistic of 0.99 ($p = 0.33$), a Mann–Whitney U test p-value of 0.25, and a permutation test p-value of 0.35, with Cohen's d of 0.32. Energy, in contrast, demonstrates almost no difference between the groups, with a t-statistic of 0.14 ($p = 0.89$) and a negligible Cohen's d of 0.04. Entropy registers a t-statistic of $-0.06$ with $p \approx 0.95$, while homogeneity ($t = -0.38$, $p > 0.70$) and dissimilarity ($t = 0.61$, $p > 0.50$) similarly fail to show significant differences. Notably, the correlation feature, despite a slightly larger effect size (Cohen's $d = -0.39$) and a t-statistic of $-1.19$, still yields non-significant p-values (0.24 for the t-test and 0.26 for the permutation test).

For the PhysMAS dataset (Figure V.47), the boxplots hint at minor visual differences. The contrast feature shows one group with a modestly higher median and a wider spread, and the correlation boxplot indicates that non-responders display greater variability compared to the more clustered responders. Nonetheless, these visual differences are not supported statistically. The contrast feature shows a t-statistic of 0.74 ($p = 0.47$) with a Cohen's $d = 0.32$. Energy has a t-statistic of $-1.15$ ($p = 0.26$, $d = -0.46$), and entropy a t-statistic of 0.38 ($p = 0.71$, $d = 0.16$). Homogeneity is nearly identical between groups ($t = 0.12$, $p = 0.91$, $d = 0.05$), while correlation ($t = -0.51$, $p = 0.61$, $d = -0.22$) and dissimilarity ($t = 0.53$, $p = 0.60$, $d = 0.22$) also fail to reach significance.

Across all datasets, although the boxplots sometimes suggest slight shifts in medians or variability, the statistical analyses consistently yield non-significant p-values and small effect sizes. This discrepancy may be attributable to the aggregation method, averaging multiple images per subject may obscure intra-subject variability, and the limited sample sizes, which constrain the statistical power.

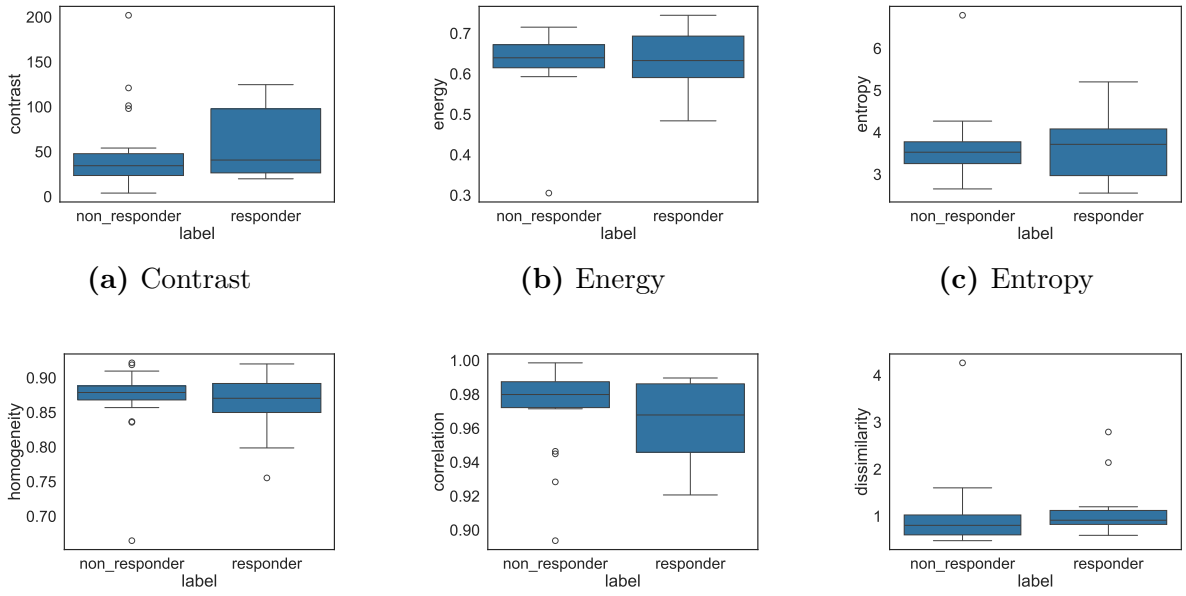



**(a)** Contrast **(b)** Energy **(c)** Entropy

**(d)** Homogeneity **(e)** Correlation **(f)** Dissimilarity

**Figure V.45:** Boxplots of GLCM features for the OSAMAS dataset illustrate the distributions of contrast, energy, and entropy (top row) as well as homogeneity, correlation, and dissimilarity (bottom row) for responders and non-responders. The overlapping distributions indicate limited discriminative power of these features for treatment response classification.

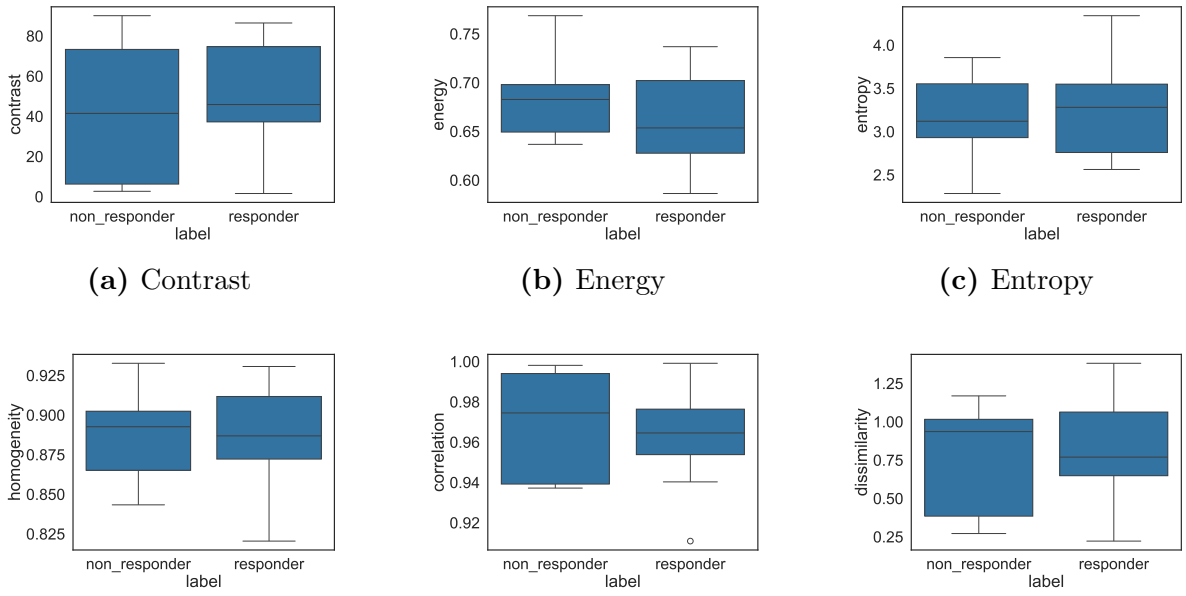
In summary, despite minor visual indications of group differences, the comprehensive statistical evaluation demonstrates that the GLCM-based texture features derived from scalogram images lack sufficient discriminative power for classifying treatment response.




**(a)** Contrast **(b)** Energy **(c)** Entropy

**(d)** Homogeneity **(e)** Correlation **(f)** Dissimilarity

**Figure V.46:** Boxplots of GLCM features for the CRC dataset illustrate the distributions of contrast, energy, and entropy (top row) as well as homogeneity, correlation, and dissimilarity (bottom row) for responders and non-responders. The overlapping distributions highlight the limited discriminative ability of these features in distinguishing treatment outcomes.




**(a)** Contrast **(b)** Energy **(c)** Entropy

**(d)** Homogeneity **(e)** Correlation **(f)** Dissimilarity

**Figure V.47:** Boxplots of GLCM features for the PhysMAS dataset illustrate the distributions of contrast, energy, and entropy (top row) as well as homogeneity, correlation, and dissimilarity (bottom row) for responders and non-responders. The overlapping distributions further underscore the limited discriminative capability of these texture measures in distinguishing treatment response.

The consistently overlapping distributions, non-significant p-values, and small effect sizes indicate that the observed differences are likely due to random variation. These findings suggest that alternative feature extraction methods or more advanced modeling approaches may be necessary to capture the subtle nuances associated with treatment outcomes.

### V.4.3 Application of Classical Machine Learning Methods for Classification of Obstructive Sleep Apnea Dataset

To evaluate our classical machine learning framework, we applied logistic regression, random forests, support vector machines, and XGBoost to time-domain and spectral features extracted from the OSAMAS, CRC, and PhysMAS datasets, see Subsection IV.6.1. We assessed the performance of each feature set individually as well as in combination. Initial experiments using stratified cross-validation yielded unsatisfactory results due to the limited sample size, which led us to adopt a leave-one-out cross-validation strategy.

We also hypothesize that the sampling frequency is critical for capturing discriminative features. Therefore, each dataset was analyzed separately, and additional experiments were conducted by combining the CRC and PhysMAS datasets, which have higher sampling frequencies than OSAMAS. Finally, all three datasets were combined for analysis. A decline in performance when including OSAMAS may further suggest that higher sampling frequencies are essential for effective signal classification.

In our classification framework, time-domain and spectral features are extracted from `.csv` files, allowing for the analysis of each feature set individually or in combination by appropriately prefixing and merging the data. The predictions are aggregated over folds to compute performance metrics, including accuracy, recall, precision, F1 score, specificity, Cohen's kappa, balanced accuracy, and ROC-AUC. Additionally, the pipeline produces diagnostic plots, such as confusion matrices, ROC curves, and precision-recall curves and incorporates hyperparameter tuning for the support vector machine via grid search.

Across all experiments, the OSAMAS dataset consistently yielded the poorest performance. In contrast, combining the CRC and PhysMAS datasets resulted in significantly better metrics than when all three datasets were merged, suggesting that incorporating OSAMAS data may dilute the models' discriminative power. The best overall performance was observed with the Random Forest model applied to the combined CRC and PhysMAS data using both temporal and spectral features, with Random Forests consistently delivering the highest recall compared to the other classifiers.

Although Random Forests outperformed the other algorithms in most cases, it should be noted that for the combined three-dataset scenario, XGBoost, when trained solely on spectral features, achieved the highest performance, with a mean accuracy of 61.8 % and a recall of 55.3 %. In most instances, the performance metrics using combined features were

nearly identical to those obtained with only spectral features, underscoring the generally lower discriminative power of time-domain features. An exception was observed with the PhysMAS dataset, where spectral features alone yielded better results than the combined feature set. We also observed that Random Forests performed particularly poorly when trained exclusively on time-domain features, almost at the level of a random classifier. Detailed performance metrics of the best performing model on combined feature set – Random Forests, are presented in Tab. V.14.

**Table V.14:** Average Classification Metrics for Random Forest on Individual Datasets, CRC and PhysMAS Combined, and All Three Datasets Combined

| Model | Acc. | Recall | $F_\beta$ Score | Prec. | $\kappa$ | Bal. Acc. | Spec. | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| OSAMAS | 39.0 | 33.3 | 32.4 | 31.6 | −0.023 | 38.4 | 43.5 | 0.338 |
| CRC | 81.1 | 60.0 | 72.0 | 90.0 | 0.586 | 77.7 | 95.5 | 0.755 |
| PhysMAS | 66.7 | 78.6 | 73.3 | 68.8 | 0.294 | 64.3 | 50.0 | 0.593 |
| CRC and PhysMAS | 73.8 | 65.5 | 70.4 | 76.0 | 0.471 | 73.4 | 81.2 | 0.775 |
| OSAMAS, CRC, PhysMAS | 51.0 | 40.4 | 43.2 | 46.3 | 0.004 | 50.2 | 60.0 | 0.502 |

**Note:** Acc. = Accuracy; Prec. = Precision; Spec. = Specificity; Bal. Acc. = Balanced Accuracy.

Next, we analyze the diagnostic plots for the Random Forest classifier applied to the combined CRC and PhysMAS dataset using the full feature set. The confusion matrix (Fig. V.48a) shows that the classifier accurately identifies most non-responders (81.2 %), while it correctly classifies only about 65.5 % of responders, indicating a bias toward predicting the majority class. Although time-domain features alone perform poorly, almost like a random classifier, their inclusion in the full feature set appears to enhance the prediction of non-responders, contributing to more balanced overall metrics. Additionally, the ROC curve (Fig. V.48c), with an AUC of approximately 0.77, demonstrates moderate overall discrimination, and the precision-recall curve (Fig. V.48b) reveals that the model's performance is sensitive to threshold variations. These behaviors are typical in small, imbalanced datasets, where overfitting to the dominant class can undermine the detection of the minority class. In such cases, increasing the dataset size or applying rebalancing techniques may help the model establish a more balanced decision boundary and improve responder detection.

To assess the impact of excluding time-domain features, we evaluated the Random Forest classifier trained exclusively on spectral features from the combined dataset. The corresponding confusion matrix (Fig. V.48g) reveals that the accuracy for non-responders decreases from approximately 81 % to 75 %, while the accuracy for responders remains around 65%. In addition, the ROC-AUC slightly declines from 0.77 to 0.75. These findings suggest that although relying solely on spectral features produces more balanced overall metrics, excluding time-domain features leads to a modest yet noticeable reduction in

non-responder prediction. In contrast, using only time-domain features and excluding spectral features significantly diminishes the model's discriminative ability, producing an average accuracy of 55.7 %, recall of 41.4 %, and specificity of 68.8 %. Given the limited dataset size and inherent class imbalance, retaining both spectral and temporal features appears essential for stronger predictive performance.

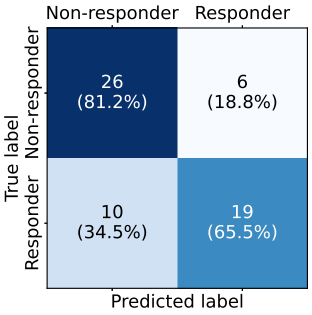
To improve the overall metrics, our next step was to refine the evaluation framework. First, we decided to dismiss the OSAMAS dataset because its inclusion degraded model performance when combined with the CRC and PhysMAS datasets. In addition, we noted that the SVM classifier achieved better results on exclusively temporal features than Random Forests. This observation motivated us to experiment with an ensemble approach that combines the strengths of both classifiers.

Ensemble models that focus on different aspects of the data can potentially yield improved performance. In our study, although the Random Forest model trained on the combined feature set performed better than other models overall, its handling of time-domain features individually was suboptimal compared to that of the SVM. Therefore, we implemented an ensemble that combined the strengths of both models, employing Random Forests for the combined features and SVM for the temporal features, using soft-voting strategy. The ensemble did not outperform the standalone Random Forest model, with the best results (obtained at a threshold of 0.44) summarized in Tab. V.15.

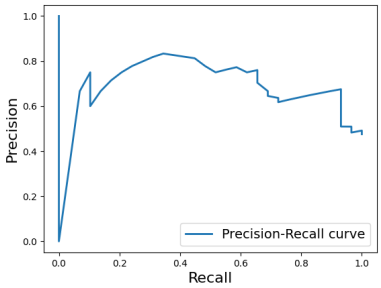
In response to the Random Forest's tendency to favor the non-responder class, we enhanced our evaluation pipeline with three key modifications. First, we experimented with various decision threshold values between 0.4 and 0.5 (using 0.1 increments) and found that lowering the threshold from the standard 0.5 to an optimal value of 0.46 significantly improved recall on the combined CRC and PhysMAS dataset. To address class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which generates synthetic minority class samples by interpolating between existing examples and their nearest neighbors, thereby enhancing the classifier's ability to learn the decision boundary.

Finally, we evaluated three configurations for the Random Forest classifier – an uncalibrated model and two calibrated models using isotonic regression and Platt scaling (see Tab. V.15). The uncalibrated model, implemented with a threshold of 0.46, achieved an accuracy of 70.5 %, recall of 72.4 %, and specificity of 68.8 %. However, its overall recall was lower than that of the Platt-scaled model, and since our application prioritizes high recall, we selected the Platt-scaled configuration as our best model.

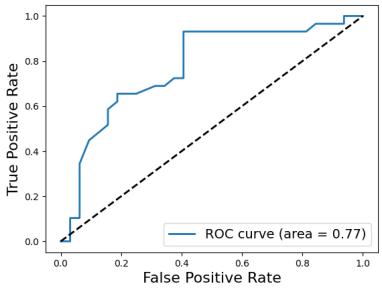
Platt scaling (Platt, 1999) maps raw classifier scores to calibrated probabilities by fitting a sigmoid function: this parametric approach is particularly robust when calibration



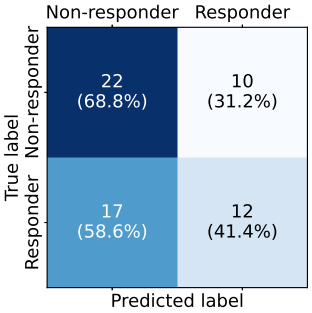

**(a)** Confusion matrix, full feature set



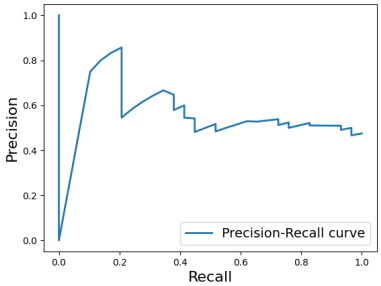

**(b)** Precision-recall curve, full feature set



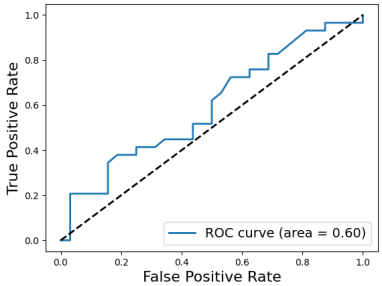

**(c)** ROC curve, full feature set



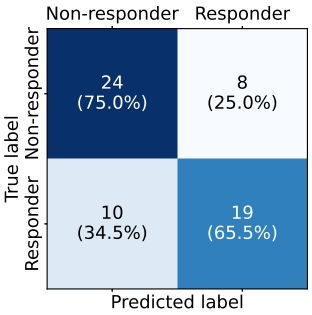

**(d)** Confusion Matrix, time features



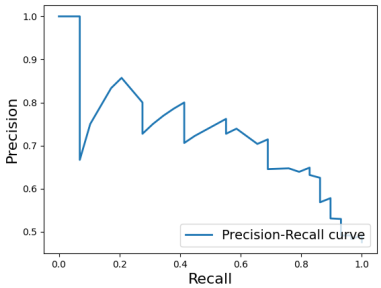

**(e)** Precision-recall curve, time features



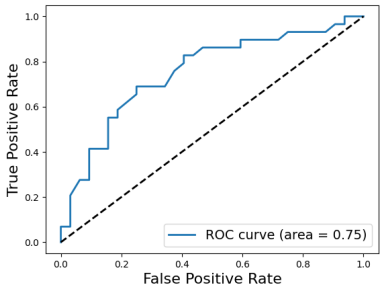

**(f)** ROC curve, time features




**(g)** Confusion matrix, spectral features




**(h)** Precision-recall curve, spectral features




**(i)** ROC curve, spectral features

**Figure V.48:** Comparison of Random Forest classification performance using: (top row) the full feature set, (middle row) time-only features, and (bottom row) spectral-only features. The confusion matrices, Precision-recall curves, and ROC curves illustrate differences in model behavior when including or excluding temporal or spectral information.

data is limited or when the relationship between scores and true probabilities closely follows a sigmoid curve. In contrast, isotonic regression, a non-parametric method often recommended for Random Forests, produced more balanced metrics (accuracy of 63.9 %, recall of 65.5 %, and specificity of 62.5 %) but ultimately performed worse than the Platt-scaled model, likely due to its tendency to overfit on small datasets (Zadrozny et al., 2002). Consequently, Platt scaling provided a better bias-variance trade-off and led to improved overall performance.

The modified Random Forest model (metrics summarized in Row 1 of Tab. V.15) demonstrates a strong capacity for detecting the responder class. This is evident from the high recall of 89.7 % and an $F_\beta$ score of 83.9 % (with $\beta = 2$), which was chosen instead of the standard F1 score to make the evaluation more sensitive to the responder class. When $\beta > 1$, recall is weighted more heavily than precision, which is particularly useful in our context where correctly identifying responders is more critical, even if it means a slight decrease in precision. As shown in the confusion matrix (Fig. V.49a), the model correctly identifies 26 out of 29 responders, although its specificity is more modest at 59.4 %, indicating that it often misclassifies non-responders as responders. This trade-off is further reflected in a precision of 66.7 %, meaning that approximately two-thirds of the predicted responders are true positives.

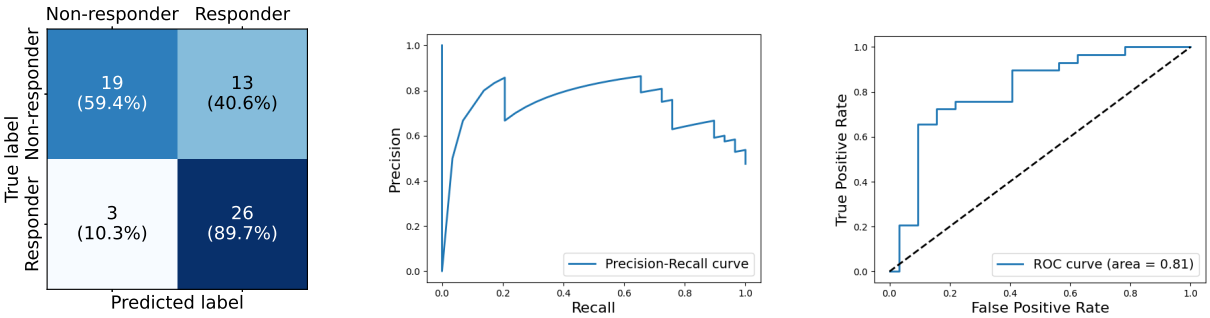
Overall accuracy of 73.8 % and the balanced accuracy of 74.5 % confirm that the model performs reasonably well across both classes. The ROC AUC of 0.81 suggests strong overall discriminative power, with the ROC curve (Fig. V.49c) showing a favorable balance between true and false positive rates as the decision threshold varies. In addition, the precision-recall curve (Fig. V.49b), with an average precision of 74.4 %, illustrates that the model can achieve high precision at certain recall levels, though precision diminishes as recall increases – a typical pattern when dealing with class imbalance.

**Table V.15:** Average Classification Metrics for Random Forest and Random Forest Ensembled with SVM on CRC and PhysMAS Combined Dataset

| Model | Acc. | Recall | $F_\beta$ | Prec. | $\kappa$ | Bal. Acc. | Spec. | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| Random Forest (Platt Scaler) | 73.8 | 89.7 | 83.9 | 66.7 | 0.483 | 74.5 | 59.4 | 0.809 |
| Random Forest (Unscaled) | 70.5 | 72.4 | 71.4 | 67.7 | 0.410 | 70.6 | 68.8 | 0.767 |
| Random Forest (Isotonic Regression) | 63.9 | 65.5 | 64.6 | 61.3 | 0.279 | 64.0 | 62.5 | 0.735 |
| Random Forest and SVM Ensemble | 50.8 | 58.6 | 56.3 | 48.6 | 0.023 | 51.2 | 43.8 | 0.439 |

**Note:** Acc. = Accuracy; Prec. = Precision; Spec. = Specificity; Bal. Acc. = Balanced Accuracy; $F_\beta$ = $F_\beta$ Score.

For the combined CRC and PhysMAS data, we performed a feature importance analysis using the Random Forest classifier on the combined time-domain and spectral feature




**(a)** Confusion matrix **(b)** Precision-recall curve **(c)** ROC curve

**Figure V.49:** Performance of the modified Random Forest classifier on the combined CRC and PhysMAS dataset. The plots display the confusion matrix, the precision-recall curve, and the ROC curve, providing a comprehensive view of the model's behavior.

set. As shown in Fig. V.50, mid-range spectral features, such as `spec_Mid2`, stand out as particularly influential for distinguishing responders from non-responders, while various time-domain metrics and other spectral features also contribute but appear less critical than the top-ranked mid-range frequencies. This suggests that although Random Forest leverages information from multiple dimensions, the model primarily relies on spectral content in the mid- to higher-frequency range.

Our tests were performed on the combined CRC and PhysMAS dataset with features extracted from `flow_data_sleep_stages.csv` (see Subsection IV.5.2). We also evaluated our best-performing model, the modified Random Forest model, on datasets containing `flow_data_REM.csv` signals – recorded only during the REM stage of sleep – and `flow_data_respiratory.csv` signals, which record abnormal breathing activity. The model performed significantly worse on `flow_data_REM.csv`, yielding results no better than a random classifier. On `flow_data_respiratory.csv`, it achieved more modest results than with `flow_data_sleep_stages.csv` (mean accuracy: 62.3 %, recall: 62.1 %, specificity: 62.5 %, and precision: 60.0 %). Although the metrics were more balanced for `flow_data_respiratory.csv`, indicating similar predictive capacity for both classes, the dataset comprising longer signals, that is measurements taken between respiratory events, exhibited stronger performance metrics.

Generally, our findings confirm that spectral features, especially those in the mid- to higher-frequency range, possess the greatest discriminative power. Prioritizing these features may yield significant performance gains, even though integrating select time-domain characteristics offers additional predictive value. Moreover, the consistently superior performance of the Random Forest model compared to other classifiers, with Logistic Regression showing the weakest results, suggests that the underlying decision boundary is non-linear.

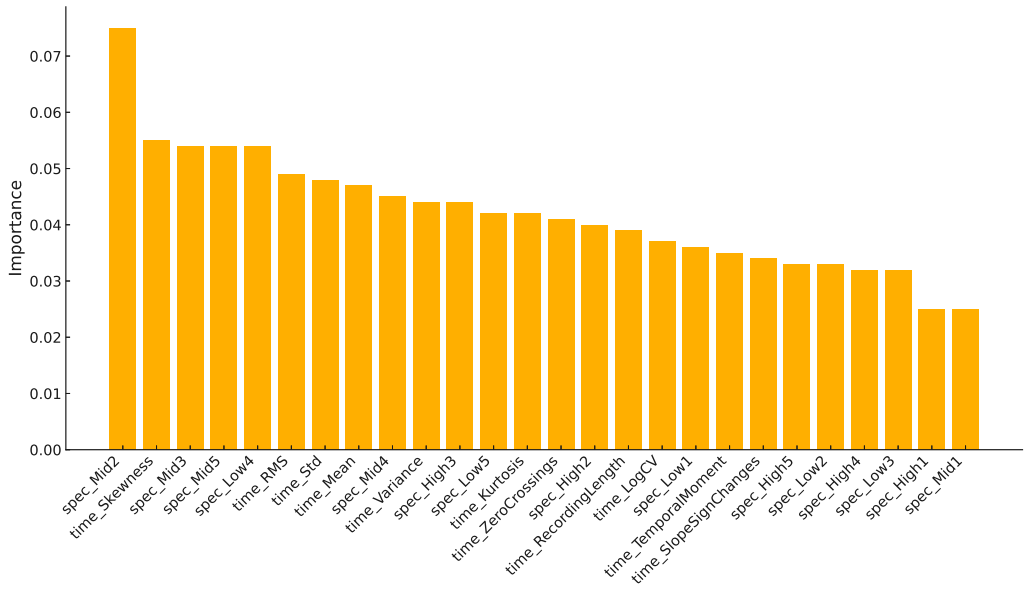



**Figure V.50:** The figure displays the relative importance of each feature as determined by the Random Forest model. The bar heights represent each feature's contribution to reducing impurity in the decision trees, normalized so that the values sum to 1. Notably, mid-range spectral features, such as `spec_Mid2`, are particularly influential.

Although our model demonstrates an improved ability to predict responders to MAS treatment, the limited dataset size increases the risk of overfitting and may compromise the generalizability of our findings to new, unseen data. Moreover, the lower specificity suggests a tendency to over-predict responders, potentially leading to a higher false-positive rate. Therefore, further validation on larger datasets is essential to confirm the model's robustness, reliability, and potential for clinical application.

### V.4.4 Testing Random Forest Model on Downsampled Signals of Obstructive Sleep Apnea Dataset

As discussed in IV.6.1, the analysis of the OSAMAS dataset is constrained by its low sampling rate, which limits the frequency range to 0–5 Hz. This restriction prevents evaluation of higher-frequency content that is present in the CRC and PhysMAS datasets, where substantial energy extends beyond 5 Hz. Furthermore, in the CRC and PhysMAS datasets, PSD curves reveal notable differences between responders and non-responders at frequencies above 10 Hz, whereas the PSDs for the OSAMAS dataset appear nearly identical for both classes. These observations suggest that mid- and high-frequency components may serve as key discriminative features for classification.

Earlier evaluations, summarized in Tab. V.14, showed that the best classification model, effective on the PhysMAS and CRC datasets, performed no better than random guessing on the OSAMAS dataset. Subsequent experiments, which involved separate assessments

of time-domain and spectral features, indicated that spectral features, particularly those from mid-range frequencies, had a critical impact on classification performance. We hypothesize, therefore, that the inability to classify the OSAMAS dataset stems from the absence of mid- and high-frequency information.

To further test this hypothesis, we conducted an experiment using the best-performing model on downsampled versions of the PhysMAS and CRC signals. These signals were reduced to a sampling rate of 10 Hz, and the resulting performance metrics are presented in Tab. V.16. If the model maintains performance on the downsampled signals comparable to that on the original high-frequency signals, it would support the argument that the mid- and high-frequency bands are essential for accurate classification.

The downsampling procedure begins by estimating the original sampling frequency from the time stamps, using the average interval between successive samples. When the original frequency exceeds the target rate, a decimation factor is computed as their ratio. The MATLAB `decimate` function is then applied to reduce the signal's sampling rate; this function employs an anti-aliasing filter to suppress frequency components above the new Nyquist limit. Concurrently, the time stamps are subsampled by retaining every $n$th value, where $n$ is the decimation factor, and the effective sampling rate is updated accordingly. This approach produces a downsampled signal that approximates one recorded natively at the lower frequency while mitigating aliasing effects.

In theory, downsampling a high-frequency signal with an ideal anti-aliasing filter preserves all information within the target frequency band, assuming the original signal is strictly band-limited below the new Nyquist limit. However, in practical applications, differences often arise. The digital anti-aliasing filter used in downsampling may differ from the analog filtering present in native low-frequency recording systems, and sensor noise along with hardware response characteristics may vary. Therefore, while downsampling can yield a signal functionally equivalent for the frequency band of interest, subtle differences in noise and overall signal characteristics may persist (Oppenheim, 1999; Proakis, 2001). It is important to acknowledge that the downsampled dataset does not exactly replicate a signal recorded at a native low sampling rate.

We first evaluated the modified best-performing model, which had previously yielded the highest recall when tested on signals at the original sampling rate. The corresponding metrics are summarized in Row 1 of Tab. V.16. A detailed examination of the confusion matrix reveals a pronounced bias towards the responder class: every non-responder instance was misclassified as a responder. This misclassification results in a specificity of 0.0 % and an overall accuracy of 45.9 %, which is below expectations for a balanced classifier. Although the responder class achieves a high recall of 96.6 %, this comes at the expense of misclassifying non-responders, as evidenced by the low precision of 46.7 %,

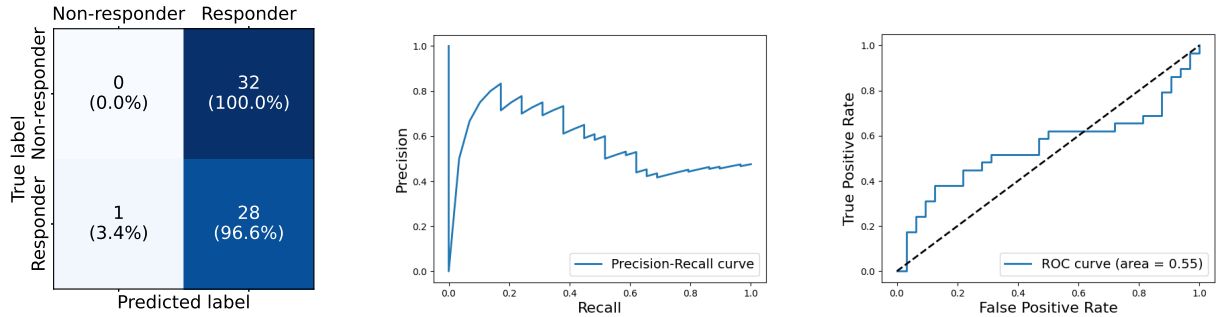
nearly half of the instances predicted as responders are actually non-responders.

Despite the high recall, the $F_\beta$ score of 79.5 % is largely driven by the model's ability to capture responder instances, without adequately recognizing non-responders. The Cohen's Kappa value of −0.033 indicates that the overall performance is effectively no better than random chance when both classes are considered. Furthermore, the balanced accuracy of 48.3 % reinforces the model's failure to meaningfully identify non-responders. From a ranking perspective, the ROC curve produces an AUC of 0.551, which is only marginally above random chance. Similarly, the precision-recall curve, with an average precision of 58.9 %, reflects the persistent imbalance in the predictions.

**Table V.16:** Average Classification Metrics for Random Forest with downsampled CRC and PhysMAS and OSAMAS Datasets

| Model | Accuracy | Recall | $F_\beta$ Score | Prec. | $\kappa$ | Balanced Accuracy | Spec. | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 45.9 | 96.6 | 79.5 | 46.7 | −0.033 | 48.3 | 0.00 | 0.551 |

**Note:** Prec. = Precision; Spec. = Specificity.




**(a)** Confusion matrix **(b)** Precision-recall curve **(c)** ROC curve

**Figure V.51:** Performance of the Random Forest classifier for the classification of downsampled CRC and PhysMAS Datasets.

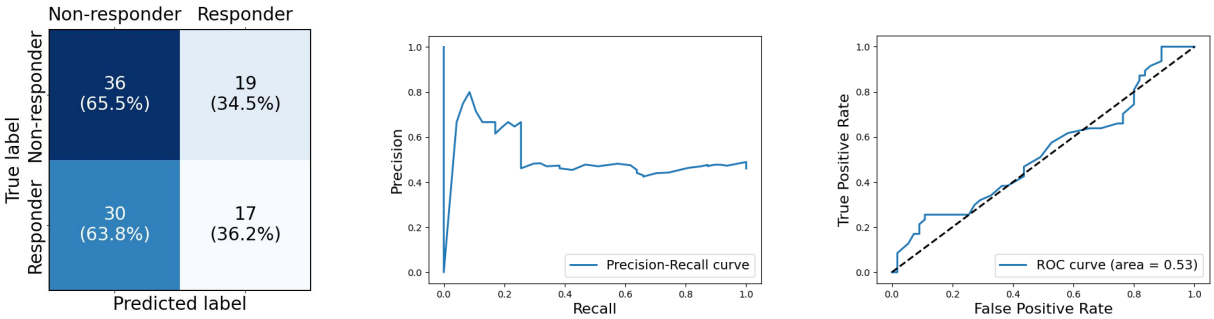
We also conducted an experiment combining the OSAMAS dataset with downsampled versions of the CRC and PhysMAS datasets. This approach created a larger sample pool while ensuring that the spectral features were confined to the same frequency bands, which should theoretically enhance the classifier's training. The evaluation of the original Random Forest model on this combined dataset is summarized in Tab. V.17, where an overall accuracy of 52.0 % was achieved. A closer look at the confusion matrix reveals that 65.5 % of non-responders were correctly identified, while 34.5 % were misclassified as responders. In contrast, only 36.2 % of responders were correctly classified, reflecting the model's difficulty in effectively identifying the responder class. This imbalance is further underscored by a recall of 36.2 % and a precision of 47.2 %.

**Table V.17:** Average Classification Metrics for Random Forest with Downsampled CRC and PhysMAS and OSAMAS Datasets

| Model | Accuracy | Recall | $F_\beta$ Score | Prec. | $\kappa$ | Balanced Accuracy | Spec. | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 52.0 | 36.2 | 41.0 | 47.2 | 0.017 | 50.8 | 65.5 | 0.527 |

**Note:** Prec. = Precision; Spec. = Specificity.

The specificity of 65.5 % suggests moderate success in detecting non-responders; however, a balanced accuracy of 50.8 % indicates that the classifier's overall performance is only marginally better than random guessing when both classes are considered. The $F_\beta$ score of 41.0 % further highlights the imbalance between precision and recall, while a near-zero Cohen's Kappa value of 0.017 emphasizes the model's weak overall agreement with the ground truth. From a ranking perspective, the ROC curve yields an AUC of 0.527, only slightly exceeding the 0.5 threshold of a random classifier, and the average precision from the precision-recall curve is 52.4 %, again reflecting modest discriminative capability. Overall, while the Random Forest classifier demonstrates marginally better-than-chance performance on some metrics, its limited ability to accurately distinguish responders from non-responders suggests that further refinement of model parameters or alternative classification strategies is necessary to achieve clinically meaningful results.




**(a)** Confusion matrix **(b)** Precision-recall curve **(c)** ROC curve

**Figure V.52:** Performance of the Random Forest classifier for the classification of combined OSAMAS downsampled CRC and PhysMAS Datasets.

Our findings suggest that reducing the sampling frequency negatively affects classification performance by removing essential spectral details, especially in mid- and high-frequency bands. This loss limits the availability of discriminative features, hindering the model's ability to effectively distinguish responders from non-responders. Thus, preserving higher-frequency content appears crucial to capturing subtle physiological differences reflected in airflow signals. Ensuring adequate sampling frequency is therefore important for developing robust classification approaches.

## Chapter VI

# Conclusion and Outlook

Routine physiological monitoring provides an abundant yet sometimes under-exploited source of diagnostic information. Previously, we argued that spectral decomposition could reveal clinically relevant patterns that are not apparent in the raw time-amplitude domain, and we posed three guiding objectives: to automate EMG classification, to predict MAS treatment efficacy in OSA, and to test whether spectral and machine learning methods are applicable across diverse physiological contexts. The experimental results presented in previous chapters support these objectives.

First, deep learning ensembles trained on automatically extracted time–frequency features were able to distinguish chloride- from sodium-channel myotonias while maintaining well-calibrated probability outputs and physiologically interpretable saliency maps. Second, classical machine learning models operating on engineered spectral features of baseline nasal-airflow signals identified responders to MAS with recalls approaching 90 %. Together, these findings indicate that spectral characteristics encode pathophysiological mechanisms in both electrical and mechanical waveforms and can be used for clinically meaningful prediction tasks.

Beyond demonstrating feasibility, the work shows how data quality and uncertainty quantification govern the translational potential of spectral analytics. It also emphasizes the value of modality-agnostic feature learning: despite differences in etiology, anatomy, and signal origin, a spectral analysis framework delivered robust performance across multiple clinical scenarios. The remainder of this chapter summarizes these results, examines their limitations, and outlines specific directions for enhancing clinical application through larger datasets, improved model calibration, and improved interpretability techniques.

# VI.1 Conclusion: Skeletal Muscle Disorders Datasets

## VI.1.1 Discussion and Implications: Skeletal Muscle Channelopathy Dataset

The experimental evidence supports the central premise that chloride- and sodium-channel myotonias leave separable patterns in iEMG. Across four deep backbones and two wavelet domains the best single model attained a balanced-accuracy of 84.4 % on validation and 74.5 % on test data, while the heterogeneous four-network ensemble reached 81.4 % ± 4.2 % std balanced-accuracy over the splits. These values, confirm H1 and H2: automatically extracted time–frequency features capture pathophysiological differences. The ensemble's reliability curve tracked the diagonal closely, yielding an ECE of 0.11 and a Brier score of 0.14, indicating that its probability outputs are reasonably well calibrated. Moreover, a selective-prediction study showed that restricting decisions to cases with confidence over 0.85 would raise accuracy to ≈ 90 %, demonstrating that the model can quantify and trade off its own uncertainty for higher precision when required.

Qualitative inspection using Grad-CAM reinforced this conclusion. Saliency maps for chloride-channel defect cases exhibited multiple broadband blobs separated by frequency-silent gaps. In contrast, sodium-channel defect cases displayed a single, continuous high-frequency ridge. This observation aligns with findings by Drost et al. (2015), who reported that an IDI longer than 30 ms is typical of chloride-channel defect-induced myotonic discharges. Simulated recordings reproduced this dichotomy: saliency for sodium-channel defects formed an elongated horizontal band spanning the entire duration of the discharge, whereas saliency for chloride-channel defects condensed into a broadband spot at burst onset. Therefore, the network likely utilizes genuine pathophysiological characteristics rather than spurious artifacts, providing partial support for hypothesis H4.

Several distinct Grad-CAM patterns identified in the simulated data merit further discussion. For the sodium-channel defect class, the saliency map emerges during stimulus-evoked spikes and extends continuously across the full myotonic burst, suggesting that the network attributes importance to both the evoked action potentials and the subsequent myotonic discharge. Conversely, chloride-channel defects exhibit a compact saliency spot focused around the discharge onset. Here, the network assigns the highest importance to the initiation phase of the myotonic discharge and emphasizes a broader frequency range compared to the sodium-channel defect class. These differences highlight distinct spectral characteristics inherent to the two ion-channel pathologies.

From a diagnostic standpoint (H3), the classifier's asymmetric error profile is noteworthy. Chloride myotonia is recognized with 90 % recall, whereas sodium defects reach only 72 %. In practice this means the tool is unlikely to miss chloride cases, but may over-call

chloride in atypical sodium recordings. Even so, an 80 % overall accuracy already offers value as an aid that can prioritize CLCN1 over SCN4A sequencing.

At present, genetic confirmation is the only accepted diagnostic standard for skeletal-muscle chloride and sodium channelopathies. It is technically straightforward yet clinically burdensome. Commercial neuromuscular panels typically cost 300-1500 USD per patient depending on payer status (Roggenbuck, 2022) and the bill can rise further when single-gene tests are repeated after inconclusive results. Interpretation is equally demanding: more than 200 pathogenic CLCN1 and about 83 SCN4A mutations are already cataloged (Brenes et al., 2021), yet statistics show that up to 42 % of multi-gene panel reports return one or more variants of uncertain significance (Johnson et al., 2024). In case of myotonias, this might be related to splice-site or copy-number changes that need bespoke functional assays (Sparber et al., 2020; Koczwara et al., 2022).

Sometimes the clinical impact of a given variant is challenging to predict due to overlapping phenotypes, complex inheritance patterns, and the occasional presence of multiple pathogenic alleles within a single individual (Suetterlin et al., 2022; Maggi et al., 2017; Vacchiano et al., 2023). As a result, the same mutation can manifest with different severity, whereas distinct mutations, whether in one gene or in both, can converge on similar clinical pictures (Wu et al., 2002). Thus, electrophysiological characterization of myotonia could significantly enhance phenotype confirmation, a goal addressed by our developed classification pipeline.

Automated EMG interpretation for skeletal-muscle chloride and sodium channelopathies has not yet been integrated into clinical pathways, making this work exploratory and primarily aimed at triage: helping clinicians decide whether to begin with targeted CLCN1 sequencing or a broader multigene panel. For this purpose, clinical usefulness depends less on overall accuracy and more on whether specific operating points achieve sufficient reliability for informed decision-making.

Based on the reported metrics, two clinically meaningful regions of operation can be defined:

- **Rule-in for chloride class.** Choose a threshold at which the model labels a case as “chloride” to maximize precision, thus minimizing false chloride calls while maintaining high recall to ensure few missed chloride cases. A practical goal for external validation would be precision $\geq$ 0.85 and recall $\geq$ 0.90 for the chloride class, and specificity $\geq$ 0.90 for sodium class at the same threshold. These thresholds can be visualized on ROC curves to illustrate trade-offs between sensitivity and false positives.
- **Abstain / defer region.** Define a confidence range in which the model issues no automatic decision and defers to standard panel testing. Within the actionable (high-

confidence) region, Cohen's $\kappa \geq 0.70$ would indicate substantial agreement, while calibration improvements should target ECE $\leq$ 0.05 and Brier $\leq$ 0.10. Reporting coverage versus confidence will show the proportion of cases in which the model provides actionable predictions.

On the current single-center dataset, the ensemble achieved metrics fall short of the above external-validation targets but demonstrate the potential to reach clinically relevant precision and recall once calibration is improved and the model is validated across multiple centers. The high-confidence subset (accuracy $\approx 90\,\%$ for confidence $> 0.85$) is particularly promising in this respect.

Because the proposed tool would serve only as a triage aid, its errors mainly affect testing logistics rather than patient safety. A false "chloride" classification could lead to ordering a single-gene CLCN1 test first, followed by a broader panel if results are negative, which adds time and cost but minimal risk. Conversely, a false "sodium" prediction typically leads directly to panel testing, which is conservative and avoids diagnostic delay.

As software influencing diagnostic decision pathways, this system falls under the regulatory definition of Software as a Medical Device. Regulatory frameworks require evidence of (1) *scientific validity* — the physiological rationale linking model features to pathology, (2) *analytical validity* — robustness to variations in acquisition and preprocessing, and (3) *clinical performance* — validated metrics in the intended population (International Medical Device Regulators Forum (IMDRF), 2017). Meeting these requirements typically involves multicentre external validation and calibration assessment in realistic clinical settings (Collins et al., 2024; STARD-AI Steering Group, 2025).

Our present study, based on data from a single center, represents an analytical validation stage. Before clinical deployment, the model would require independent external testing, prospective evaluation, and usability assessment to confirm its performance and safety within clinical workflows.

An inexpensive and routinely performed iEMG screening analysis capable of differentiating chloride- from sodium-channel myotonias would improve clinical diagnostic efficiency and accuracy. By pre-selecting cases prior to genetic analysis, this electrophysiological classifier could reduce testing costs, lower the downstream burden of variants of uncertain significance, and shorten the diagnostic process. Even when its predictions are imperfect, it delivers tangible clinical and economic value and represents a pragmatic step toward integrating electrophysiological insights with molecular diagnostics.

### VI.1.2 Discussion of Fibrillation Potentials Dataset: Analysis of the Algorithm's Universal Applicability

The fibrillation potentials dataset provides an independent and physiologically distinct benchmark against which to gauge the scope of our pipeline applied to channelopathy type classification. On this task the best single model attained a test balanced accuracy of 86.2 % and a ROC-AUC of 0.94, while the eight-member ensemble converged at 86.9 % balanced-accuracy and 0.93 ROC-AUC, markedly higher than the 81.4 % achieved on the ion-channel defect type discrimination experiment. Cross-validated dispersion narrowed to ±8.2 % on accuracy and ±4.4 % on ROC-AUC, and only one split fell below 75 % balanced-accuracy, indicating stable performance. Such parity, or improvement, without architecture or hyper-parameter changes confirms that automatically extracted time-frequency features and transfer learning scheme can be applied beyond ion-channel myotonia problem, supporting H2.

Physiological plausibility is maintained for this signal type, with class-specific saliency patterns consistent with classical EMG descriptions. Correctly classified fibrillation potentials are characterized by spectrally broad, temporally brief bursts of energy spanning a wide frequency range, visually represented as vertically extensive hotspots in Grad-CAM analyses. In contrast, correctly identified voluntary contractions exhibit sustained, narrow-band activity concentrated primarily in higher frequency ranges, appearing as elongated horizontal ribbons within scalograms.

Fibrillation potentials physiologically arise from spontaneous, asynchronous, and brief twitches of individual denervated muscle fibers. Their broad spectral profile and energy distribution across multiple frequencies reflect the sharp, irregular, and variable waveform patterns generated by isolated, unsynchronized muscle fiber discharges. Each isolated discharge typically produces complex signals covering diverse frequency components due to its distinct morphology and lack of temporal synchronization.

Voluntary contractions, however, result from coordinated activation of multiple MUs, each innervating numerous muscle fibers. The observed concentration of saliency into a narrower high-frequency band likely indicates that synchronized or near-synchronized firing of multiple motor MUs generates a structured and temporally sustained signal, dominated by characteristic higher frequency ranges. This structured pattern may correspond to typical motor neuron firing rates and the morphology of compound MU action potentials.

Misclassifications events provide further insight into the model's decision boundaries. A fibrillation potential misclassified as voluntary contraction due to an extended temporal saliency region might represent instances where multiple denervated fibers discharge in rapid succession or grouped manner. Such repetitive activity could mimic the sustained temporal characteristics of voluntary contractions despite retaining irregular spectral

content. Conversely, a voluntary contraction signal misclassified as fibrillation might occur if the recording captures a very brief, isolated burst of activity from a small number of MUs, or perhaps a superimposed artifact that mimics the spectral signature of a fibrillation. These distinct saliency patterns confirm the network's reliance on genuine pathophysiological cues rather than dataset artifacts, thus supporting H1.

Ensemble probabilities exhibit an expected-calibration error of $0.10 \pm 0.06$ and a Brier score of $0.11 \pm 0.03$; discarding predictions with confidence below 0.85 raises accuracy to $> 95\,\%$. Reliability curves therefore show only mild over-confidence in the mid-probability range: an outcome amenable to post-hoc temperature scaling. Selective-prediction plots demonstrate that the model recognizes its own ambiguity: on the most challenging split, accuracy exceeds $90\,\%$ once low-confidence cases are deferred.

Taken together, these findings validate the proposed universal framework. The pipeline not only detects fibrillation potentials at potentially clinically useful level but also maintains calibrated uncertainty estimates and physiologically interpretable decision logic, extending its utility from channelopathy sub-typing to spontaneous-activity detection. Residual split-to-split variance and modest calibration drift might point to data quality and quantity factors as the dominant constraints rather than model capacity, indicating that incremental gains are more likely to arise from extending the dataset, more advanced data augmentation and uncertainty-aware loss functions than from radical architectural redesign. Within those limits, the evidence shows that a single deep-learning infrastructure can accommodate heterogeneous neuromuscular phenomena, advancing the goal of a unified, automatic EMG diagnostic assistant.

### VI.1.3 Limitations and Future Research Directions

The present study demonstrates that a deep learning pipeline exploiting spectral feature extraction can classify iEMG data with high accuracy and well-calibrated probability estimates. However, three primary limitations temper this success and outline directions for further development.

**Data availability and representativeness.** The dataset used in this study remains modest relative to the requirements of contemporary deep learning models. The clustering of false-positive chloride predictions around sodium-channel defect samples indicates the classifier's sensitivity to rare phenotypes and signal acquisition artifacts. Moreover, the band-limited nature of the synthetic data ($< 1$ kHz) limits its applicability in data augmentation for training. Future improvements require larger and more diverse clinical datasets and enhanced simulations that closely mimic clinical recordings, including realistic high-frequency content.

**Model calibration and uncertainty quantification.** Expected calibration error

varied from 0.04 to 0.14 across cross-validation splits, mirroring fluctuations in class balance and noise content. While selective prediction already allows accuracy to exceed 90 % at high confidence thresholds, the model can still be improved to ensure its clinical reliability. Approaches such as temperature scaling, deep-ensemble distillation, and Bayesian last-layer adaptation provide principled means to improve calibration without incurring significant computational overhead.

**Interpretability and physiological insight.** Saliency mapping using Grad-CAM consistently highlighted potentially physiologically plausible regions aligned with classical descriptions. Nevertheless, Grad-CAM remains heuristic and can obscure complex feature interactions. Methods such as layer-wise relevance propagation, concept activation vectors, or counterfactual analyses could rigorously verify whether the model encodes specific spectral characteristics, thus supporting the identification of reliable spectral biomarkers.

Building on these observations, future work will pursue three complementary directions. First, data sharing and biophysics-aware augmentation will enlarge and re-balance the training pool, addressing the class imbalance and general data scarcity. Second, incorporating context-aware attention mechanisms, dynamic-range normalization, and noise-sensitive loss functions will strengthen model robustness against burst dominance and recording artifacts. Third, coupling calibrated ensemble predictions with clinically relevant decision thresholds, alongside prospective multi-center validation studies, will transition the classifier from a research prototype toward a clinically deployable diagnostic tool.

## VI.2 Conclusion: Obstructive Sleep Apnea Dataset

The primary aim of our study was to test the central hypothesis that baseline respiratory airflow signals contain inherent features predictive of patient response to MAS therapy in OSA. Our underlying premise was that both time-domain and spectral characteristics of the airflow, particularly morphological features of the respiratory signal, encode clinically relevant anatomical and mechanical information pertaining to upper airway patency. Thus, these signal features could potentially serve as biomarkers capable of predicting therapeutic outcomes.

To test this hypothesis, we analyzed airflow data from multiple polysomnography datasets: the OSAMAS dataset, recorded at a relatively low sampling frequency (10 Hz), and the CRC and PhysMAS datasets, captured at higher frequencies (up to 250 Hz). We standardized our preprocessing pipeline across all datasets, segmenting respiratory recordings into fixed-length epochs. For each epoch, we performed wavelet transform and obtained scalograms. From full-length signals, we extracted comprehensive features, including conventional time-domain metrics and time–frequency characteristics derived through spectral analysis methods such as PSD. This standardization ensured consistent and meaningful comparisons across datasets.

We explored the application of transfer learning techniques on scalogram image representations of airflow signals, using various deep learning architectures such as adaptive pooling methods, feature fusion strategies, fine-tuning of pre-trained ImageNet models, and custom CNN-RNN combinations. These image-based approaches exhibited limited generalization performance. Evaluation metrics consistently showed accuracy near or below chance levels, poor recall rates for the responder class, and negative Cohen's Kappa values. These findings suggested that the visual representations obtained from scalograms did not clearly differentiate responders from non-responders. Supporting this conclusion, statistical texture analyses of GLCM features also indicated insignificant differences, as the texture metrics showed considerable overlap between the responder and non-responder groups.

In contrast, classical machine learning algorithms applied directly to the extracted numerical time-domain and spectral features demonstrated more promising results. In particular, Random Forest classifiers achieved higher overall accuracy and significantly improved recall for responders when trained on datasets with higher sampling rates. Spectral features, particularly those in mid- and high-frequency bands, emerged as more discriminative compared to their time-domain counterparts. Furthermore, experiments integrating datasets indicated that inclusion of signals with low sampling rate tended to diminish predictive performance. Complementary downsampling experiments reinforced this observation; specifically, reducing the higher-frequency CRC and PhysMAS datasets to match the 10 Hz frequency of OSAMAS markedly deteriorated model performance, resulting in a near-complete loss of capability to identify non-responders.

### VI.2.1 Discussion and Implications

The findings of our study provide evidence that baseline airflow signals contain predictive information, particularly within the spectral domain that is associated with MAS treatment response. In particular, the performance of classifiers relying on spectral features extracted from high-sampling rate data shows that characteristics of frequency content are critical for distinguishing responders from non-responders.

Our best-performing Random Forest model with modified threshold, trained on a combination of the PhysMAS and CRC datasets, attained a recall of approximately 0.9, indicating that it successfully identified around 90 % of responders. This model achieved a Cohen's Kappa value of about 0.48, reflecting moderate agreement between predictions and actual classifications beyond mere chance. In comparison, a baseline model trained exclusively on the CRC dataset achieved a higher Cohen's Kappa (0.59), indicating a slightly stronger overall consistency; however, its recall was significantly lower at approximately 0.6, despite a high specificity of around 0.96. This discrepancy can be attributed primarily to pronounced class imbalance in the CRC dataset, which contained

notably fewer responder cases (15 responders versus 22 non-responders). The classifier predominantly learned characteristics of the non-responder group, diminishing its ability to accurately recognize responders. Combining the CRC dataset with the PhysMAS dataset, which had a greater proportion of responders, mitigated this imbalance and substantially improved the recall rate.

The performance of classical machine learning algorithms was also significantly influenced by variations in feature distributions arising from the different sampling frequencies among datasets. The definition and alignment of frequency bands – low, mid, and high, are inherently linked to the sampling frequency of the recorded signals. Thus, the alignment of mid-frequency bands between the CRC and PhysMAS datasets facilitated superior classification performance when these datasets were merged. Conversely, the inclusion of the OSAMAS dataset, constrained to a narrower frequency range (0–5 Hz) and thus lacking essential mid- and high-frequency content, markedly reduced the overall discriminative power. Notably, the baseline model performed optimally when trained on the CRC dataset individually (see Tab. V.14). The relatively poorer individual performance on the PhysMAS dataset, despite its higher sampling rate, likely stems from its smaller sample size (24 patients), showing that both sample size and sampling frequency jointly influence classifier performance.

Regarding image-based deep learning, while transfer learning on scalogram images was conceptually appealing, our experimental outcomes indicate that this approach is not well-suited to the current problem. Converting the airflow signals into an image format and aggregating epoch-level scalograms might tend to obscure the fine temporal dynamics that are critical for effective classification. Moreover, this method appears to suffer from the curse of dimensionality; the feature-rich overnight recordings demand a large volume of data for deep neural networks to learn distinctive features and generalize successfully. For comparison, our previous work on classifying skeletal muscle channelopathies, based on 30-second EMG recordings and approximately 500 samples, benefited from a substantially larger dataset. In the present study, however, the total of 102 patient airflow measurements is likely insufficient to fully leverage deep learning methods, leading to overfitting and poor generalization.

Our results indicate that classical machine learning models that use spectral features derived from high-sampling-rate airflow data exhibit moderate yet meaningful discriminative power. Differences in the frequency characteristics of baseline airflow signals can predict MAS treatment outcomes. However, these models are sensitive to dataset characteristics, such as class imbalances and variations in sampling frequencies. While high recall rates in some models suggest potential clinical utility for identifying responders, moderate Cohen's Kappa values and occasional imbalances in specificity underscore the need for significant improvements before a robust, clinically viable predictive tool can be developed.

### VI.2.2 Signal Quality and Its Importance for Classification

Our analysis suggests that incorporating mid- and high-frequency components is important for the classification of airflow signals. This conclusion is supported by the notable decline in performance when the PhysMAS and CRC datasets were downsampled to 10 Hz and by the generally low metrics obtained from the OSAMAS dataset, which lacks frequency content above 5 Hz. These findings support our hypothesis that datasets limited in mid- and high-frequency information result in reduced classification performance compared to datasets with a wider spectral range.

In Section IV.6.1, we discussed that differences in PSDs between responders and non-responders become apparent at frequencies above 10 Hz in the CRC and PhysMAS datasets. However, the OSAMAS dataset, limited to frequencies below 5 Hz, does not exhibit these PSD differences, further highlighting the potential importance of mid- and high-frequency content in distinguishing between the groups.

To further investigate the impact of sampling frequency, we compared two signal segments from the OSAMAS dataset: one segment from OSAMAS_012 recorded at 256 Hz and another from OSAMAS_029 recorded at 10 Hz. Both segments included a single hypopnea event with an additional 30 seconds of baseline data before and after the event. Analysis using PSDs, spectrograms, and scalograms, see Subsection IV.6.1, showed that the higher-resolution scalogram from OSAMAS_012 captured a transient shift in spectral power into higher-frequency regions, which was not visible in the lower-resolution OSAMAS_029 data.

Additional support for this observation came from experiments involving the downsampling of the PhysMAS and CRC datasets to 10 Hz, followed by classification using our best-performing algorithm. As shown in Tab. V.17, these experiments yielded near-zero Cohen's Kappa values, suggesting systematic misclassification, likely due to the loss of high-frequency information.

Lastly, feature importance analysis of our top-performing model further emphasized the relevance of mid-frequency components, consistently showing their higher discriminative capability (Fig. V.50). These results indicate the necessity of maintaining sufficiently high sampling rates to capture subtle, clinically meaningful characteristics in airflow signals.

### VI.2.3 Limitations and Future Research Directions

Although the results obtained in this study are promising, several limitations must be acknowledged and addressed before the developed models can be considered for clinical implementation. One primary limitation is the relatively small patient cohort, combined with variability in recording durations, sampling rates, and signal quality. Such variability

presents challenges to the generalizability and robustness of the developed classifiers. Future studies should therefore involve larger and more balanced patient cohorts to rigorously validate and extend these initial findings.

Additionally, the physiological distinctions between MAS responders and non-responders appear to be subtle. This subtlety indicates that current feature extraction methods may benefit from integrating multiple data modalities. Incorporating additional clinical parameters or features extracted from other PSG channels, such as chin EMG, could potentially enhance classification accuracy. Furthermore, our findings emphasize the significance of high-frequency spectral information in airflow signals. Therefore, future research should prioritize data collection strategies that ensure acquisition at sufficiently high sampling frequencies.

In conclusion, while the current study does not yet present a fully validated predictive model ready for immediate clinical use, it provides a comprehensive analysis of the spectral features of airflow signals in OSA patients undergoing MAS therapy. The findings support the central hypothesis, demonstrating that baseline respiratory signals contain predictive features primarily within mid- and high-frequency spectral bands. Furthermore, the superior performance observed with classical machine learning approaches compared to image-based deep learning methods highlights the critical role of feature engineering and high-quality data. Future research should focus on expanding datasets, refining feature extraction methods, and exploring domain-specific adaptations to further improve predictive accuracy and clinical applicability.

# References


Adrian, R. and Bryant, S. "On the Repetitive Discharge in Myotonic Muscle Fibres", The Journal of Physiology, Vol. 240 (2), pp. 505–515, 1974. [citation referenced on p. 13].

Albuquerque, E. and Thesleff, S. "A Comparative Study of Membrane Properties of Innervated and Chronically Denervated Fast and Slow Skeletal Muscles of the Rat", Acta Physiologica Scandinavica, Vol. 73 (4), pp. 471–480, 1968. [citation referenced on p. 16].

Almers, W. "Potassium Conductance Changes in Skeletal Muscle and the Potassium Concentration in the Transverse Tubules", The Journal of Physiology, Vol. 225 (1), pp. 33–56, 1972. [citation referenced on p. 13].

Project "AASM Clarifies Hypopnea Scoring Criteria". 2013. [citation referenced on p. 23].

Project "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 3". 2023. [citation referenced on p. 23].

American Academy of Sleep Medicine, N. T. T. F. "Best Clinical Practices for the Sleep Center Adjustment of Noninvasive Positive Pressure Ventilation (NPPV) in Stable Chronic Alveolar Hypoventilation Syndromes", Journal of Clinical Sleep Medicine, Vol. 6 (5), pp. 491–509, 2010. [citation referenced on p. 28].

American Academy of Sleep Medicine, P. A. P. T. T. F. "Clinical Guidelines for the Manual Titration of Positive Airway Pressure in Patients with Obstructive Sleep Apnea", Journal of Clinical Sleep Medicine, Vol. 4 (2), pp. 157–171, 2008. [citation referenced on p. 28].

Azarbarzin, A., Marques, M., Sands, S. A., Beeck, S. O., Genta, P. R., Taranto-Montemurro, L., Melo, C. M., Messineo, L., Vanderveken, O. M., and White, D. P. "Predicting Epiglottic Collapse in Patients with Obstructive Sleep Apnoea", European Respiratory Journal, Vol. 50 (3), 2017. [citation referenced on p. 28].

Bamagoos, A. A., Cistulli, P. A., Sutherland, K., Madronio, M., Eckert, D. J., Hess, L., Edwards, B. A., Wellman, A., and Sands, S. A. "Polysomnographic Endotyping to Select Patients with Obstructive Sleep Apnea for Oral Appliances", Annals of the American Thoracic Society, Vol. 16 (11), pp. 1422–1431, 2019. [citation referenced on p. 8, 26].

Barkhaus, P. E. and Nandedkar, S. D. "Atypical Fibrillation and Fasciculation Potentials: An Exercise in Waveform Identification and Analysis", Muscle & Nerve, Vol. 63 (5), pp. 657–660, 2021. [citation referenced on p. 39, 40].

Basyuni, S., Barabas, M., and Quinnell, T. "An Update on Mandibular Advancement Devices for the Treatment of Obstructive Sleep Apnoea Hypopnoea Syndrome", Journal of Thoracic Disease, Vol. 10 (Suppl 1), S48, 2018. [citation referenced on p. 25].

Beeck, S. O., Vena, D., Mann, D., Azarbarzin, A., Huyett, P., Van de Perck, E., Gell, L. K., Alex, R. M., Dieltjens, M., and Willemen, M. "Polysomnographic Airflow Shapes

and Site of Collapse During Drug-induced Sleep Endoscopy", European Respiratory Journal, Vol. 63 (6), 2024. [citation referenced on p. 28, 29].

Benmalek, E., Elmhamdi, J., and Jilbab, A. "ECG Scalogram Classification with CNN Micro-architectures", Research on Biomedical Engineering, pp. 1–11, 2022. [citation referenced on p. 4].

Berry, R. B., Budhiraja, R., Gottlieb, D. J., and al. "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events", Journal of Clinical Sleep Medicine, Vol. 8 (5), pp. 597–619, 2012. DOI: 10.5664/jcsm.2172, citation referenced on p. 23].

Bjornsdottir, E., Keenan, B. T., Eysteinsdottir, B., Arnardottir, E. S., Janson, C., Gislason, T., Sigurdsson, J. F., Kuna, S. T., Pack, A. I., and Benediktsdottir, B. "Quality of Life Among Untreated Sleep Apnea Patients Compared with the General Population and Changes After Treatment with Positive Airway Pressure", Journal of Sleep Research, Vol. 24 (3), pp. 328–338, 2015. [citation referenced on p. 19].

Breiman, L. "Random Forests", Machine Learning, 45, pp. 5–32, 2001. [citation referenced on p. 110].

Breiman, L. and Spector, P. "Submodel Selection and Evaluation in Regression. The X-random Case", International Statistical Review/revue Internationale De Statistique, pp. 291–319, 1992. [citation referenced on p. 100].

Brenes, O., Barbieri, R., Vásquez, M., Vindas-Smith, R., Roig, J., Romero, A., Valle, G. d., Bermúdez-Guzmán, L., Bertelli, S., and Pusch, M. "Functional and Structural Characterization of ClC-1 and Nav1. 4 Channels Resulting From CLCN1 and SCN4A Mutations Identified Alone and Coexisting in Myotonic Patients", Cells, Vol. 10 (2), p. 374, 2021. [citation referenced on p. 189].

Bretag, A. H. "Muscle Chloride Channels." Physiological Reviews, Vol. 67 (2), pp. 618–724, 1987. [citation referenced on p. 12, 16].

Brier, G. W. "Verification of Forecasts Expressed in Terms of Probability", Monthly Weather Review, Vol. 78 (1), pp. 1–3, 1950. [citation referenced on p. 142].

Brown, E. C., Cheng, S., McKenzie, D. K., Butler, J. E., Gandevia, S. C., and Bilston, L. E. "Tongue and Lateral Upper Airway Movement with Mandibular Advancement", Sleep, Vol. 36 (3), pp. 397–404, 2013. [citation referenced on p. 26].

Brown, E. C., Jugé, L., Knapman, F. L., Burke, P. G., Ngiam, J., Sutherland, K., Butler, J. E., Eckert, D. J., Cistulli, P. A., and Bilston, L. E. "Mandibular Advancement Splint Response is Associated with the Pterygomandibular Raphe", Sleep, Vol. 44 (4), zsaa222, 2021. [citation referenced on p. 8, 41].

Buchthal, F. and Rosenfalck, P. "Spontaneous Electrical Activity of Human Muscle", Electroencephalography and Clinical Neurophysiology, Vol. 20 (4), pp. 321–336, 1966. [citation referenced on p. 16, 17, 39].

Cannon, S. C. "Ion-channel Defects and Aberrant Excitability in Myotonia and Periodic Paralysis", Trends in Neurosciences, Vol. 19 (1), pp. 3–10, 1996a. [citation referenced on p. 18].

Cannon, S. C. "Slow Inactivation of Sodium Channels: More Than Just a Laboratory Curiosity", Biophysical Journal, Vol. 71 (1), p. 5, 1996b. [citation referenced on p. 4].

Cannon, S. C. "From Mutation to Myotonia in Sodium Channel Disorders", Neuromuscular Disorders, Vol. 7 (4), pp. 241–249, 1997. [citation referenced on p. 14].

Cannon, S. C. "Channelopathies of Skeletal Muscle Excitability", Comprehensive Physiology, Vol. 5 (2), p. 761, 2015. [citation referenced on p. 2, 11, 37, 38].

Cannon, S. C., Brown, R. H., and Corey, D. P. "Theoretical Reconstruction of Myotonia and Paralysis Caused by Incomplete Inactivation of Sodium Channels", Biophysical Journal, Vol. 65 (1), pp. 270–288, 1993. [citation referenced on p. 4, 14, 18].

Caruana, R. "Multitask Learning", Machine Learning, 28, pp. 41–75, 1997. [citation referenced on p. 119].

Cerutti, S. and Marchesi, C. "Advanced Methods of Biomedical Signal Processing", John Wiley & Sons, 2011. [[citation referenced on p. 48, 49, 60, 63, 64].

Champeney, D. C. "A Handbook of Fourier Theorems", Cambridge University Press, 1987. [[citation referenced on p. 61].

Chan, A. S., Sutherland, K., and Cistulli, P. A. "Mandibular Advancement Splints for the Treatment of Obstructive Sleep Apnea", Expert Review of Respiratory Medicine, Vol. 14 (1), pp. 81–88, 2020. [citation referenced on p. 7, 30, 44].

Chapelle, O., Weston, J., Bottou, L., and Vapnik, V. "Vicinal Risk Minimization", Advances in Neural Information Processing Systems, 13, 2000. [citation referenced on p. 123].

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 16, pp. 321–357, 2002. [citation referenced on p. 179].

Chen, H., Eckert, D. J., Stelt, P. F., Guo, J., Ge, S., Emami, E., Almeida, F. R., and Huynh, N. T. "Phenotypes of Responders to Mandibular Advancement Device Therapy in Obstructive Sleep Apnea Patients: A Systematic Review and Meta-analysis", Sleep Medicine Reviews, 49, p. 101229, 2020. [citation referenced on p. 7, 27, 30].

Chen, T. and Guestrin, C. "Xgboost: A Scalable Tree Boosting System", pp. 785–794, 2016. [citation referenced on p. 111].

Cistulli, P. A., Gotsopoulos, H., Marklund, M., and Lowe, A. A. "Treatment of Snoring and Obstructive Sleep Apnea with Mandibular Repositioning Appliances", Sleep Medicine Reviews, Vol. 8 (6), pp. 443–457, 2004. [citation referenced on p. 27].

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., and TRIPOD+AI Group, "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods", BMJ, 385, e078378, 2024. DOI: 10.1136/bmj-2023-078378, citation referenced on p. 190].

Conrad, B., Sindermann, F., and Prochazka, V. "Interval Analysis of Repetitive Denervation Potentials of Human Skeletal Muscle", Journal of Neurology, Neurosurgery & Psychiatry, Vol. 35 (6), pp. 834–840, 1972. [citation referenced on p. 17].

Cortes, C. and Vapnik, V. "Support-vector Networks", Machine Learning, 20, pp. 273–297, 1995. [citation referenced on p. 110].

Davalos, L., Arya, K., and Kushlaf, H. "Abnormal Spontaneous Electromyographic Activity", 2018. [citation referenced on p. 5].

De Moortel, I., Munday, S., and Hood, A. W. "Wavelet Analysis: the Effect of Varying Basic Wavelet Parameters", Solar Physics, Vol. 222 (2), pp. 203–228, 2004. [citation referenced on p. 65, 66].

Dempsey, J. A., Veasey, S. C., Morgan, B. J., and O'Donnell, C. P. "Pathophysiology of Sleep Apnea", Physiological Reviews, Vol. 90 (1), pp. 47–112, 2010. [citation referenced on p. 19, 21, 22, 45].

DeVries, T. and Taylor, G. W. "Improved Regularization of Convolutional Neural Networks with Cutout", ArXiv Preprint ArXiv:1708.04552, 2017. [citation referenced on p. 125].

Deymeer, F., Çakirkaya, S., Serdaroğlu, P., Schleithoff, L., Lehmann-Horn, F., Rüdel, R., and Özdemir, C. "Transient Weakness and Compound Muscle Action Potential Decrement in Myotonia Congenita", Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, Vol. 21 (10), pp. 1334–1337, 1998. [citation referenced on p. 13].

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. "Decaf: A Deep Convolutional Activation Feature for Generic Visual Recognition", pp. 647–655, 2014. [citation referenced on p. 120].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. "An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale", ArXiv Preprint ArXiv:2010.11929, 2020. [citation referenced on p. 117, 118].

Drost, G., Stunnenberg, B. C., Trip, J., Borm, G., McGill, K. C., Ginjaar, I. H., Kooi, A. W., Zwarts, M. J., Engelen, B. G., Faber, C. G., Stegeman, D. F., and Lateva, Z. "Myotonic Discharges Discriminate Chloride From Sodium Muscle Channelopathies", Neuromuscular Disorders, Vol. 25 (1), pp. 73–80, 2015. [citation referenced on p. 2, 4, 5, 32, 33, 37, 38, 147, 148, 188].

Eberstein, A. and Goodgold, J. "Slow and Fast Twitch Fibers in Human Skeletal Muscle", American Journal of Physiology-Legacy Content, Vol. 215 (3), pp. 535–541, 1968. [citation referenced on p. 17].

Eckert, D. J., White, D. P., Jordan, A. S., Malhotra, A., and Wellman, A. "Defining Phenotypic Causes of Obstructive Sleep Apnea: Identification of Novel Therapeutic Targets", American Journal of Respiratory and Critical Care Medicine, Vol. 188 (8), pp. 996–1004, 2013. DOI: 10.1164/rccm.201303-0448oc, citation referenced on p. 22].

Edwards, B. A., Andara, C., Landry, S., Sands, S. A., Joosten, S. A., Owens, R. L., White, D. P., Hamilton, G. S., and Wellman, A. "Upper-airway Collapsibility and Loop Gain Predict the Response to Oral Appliance Therapy in Patients with Obstructive Sleep Apnea", American Journal of Respiratory and Critical Care Medicine, Vol. 194 (11), pp. 1413–1422, 2016. [citation referenced on p. 8, 26].

Fawcett, T. "An Introduction to ROC Analysis", Pattern Recognition Letters, Vol. 27 (8), pp. 861–874, 2006. [citation referenced on p. 102].

Ferrante, M. A. "Comprehensive Electromyography: with Clinical Correlations and Case Studies", Cambridge University Press, 2018. [[citation referenced on p. 17, 19].

Finnsson, E., Arnardóttir, E., Cheng, W.-J., Alex, R. M., Sigmarsdóttir, B., Helgason, S., Hang, L.-W., Ágústsson, J. S., Wellman, A., and Sands, S. A. "Sleep Apnea Endotypes: From the Physiological Laboratory to Scalable Polysomnographic Measures", Frontiers in Sleep, 2, p. 1188052, 2023. [citation referenced on p. 22].

Fournier, E., Arzel, M., Sternberg, D., Vicart, S., Laforet, P., Eymard, B., Willer, J.-C., Tabti, N., and Fontaine, B. "Electromyography Guides Toward Subgroups of Mutations in Muscle Channelopathies", Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, Vol. 56 (5), pp. 650–661, 2004. [citation referenced on p. 3, 5, 37].

Fournier, E. and Tabti, N. "Clinical Electrophysiology of Muscle Diseases and Episodic Muscle Disorders", Handbook of Clinical Neurology, 161, pp. 269–280, 2019. [citation referenced on p. 4].

Fournier, E., Viala, K., Gervais, H., Sternberg, D., Arzel-Hézode, M., Laforêt, P., Eymard, B., Tabti, N., Willer, J.-C., and Vial, C. "Cold Extends Electromyography Distinction Between Ion Channel Mutations Causing Myotonia", Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, Vol. 60 (3), pp. 356–365, 2006. [citation referenced on p. 4, 14, 19, 37, 38].

Fugal, D. L. "Conceptual Wavelets in Digital Signal Processing: an In-depth, Practical Approach for the Non-mathematician", (No Title), 2009. [citation referenced on p. 63].

Fuglsang-Frederiksen, A. "The Role of Different EMG Methods in Evaluating Myopathy", Clinical Neurophysiology, Vol. 117 (6), pp. 1173–1189, 2006. [citation referenced on p. 37, 39].

Genta, P. R., Kaminska, M., and Edwards, B. A. "The Importance of Mask Selection on Continuous Positive Airway Pressure Outcomes for Obstructive Sleep Apnea: An Official American Thoracic Society Workshop Report", Annals of the American Thoracic Society, Vol. 17 (10), pp. 1177–1185, 2020. DOI: 10.1513/annalsats.202007-864st, citation referenced on p. 25].

Genta, P. R., Sands, S. A., Butler, J. P., Loring, S. H., Katz, E. S., Demko, B. G., Kezirian, E. J., White, D. P., and Wellman, A. "Airflow Shape is Associated with the Pharyngeal Structure Causing OSA", Chest, Vol. 152 (3), pp. 537–546, 2017. [citation referenced on p. 28, 29].

Goupillaud, P., Grossmann, A., and Morlet, J. "Cycle-octave and Related Transforms in Seismic Signal Analysis", Geoexploration, Vol. 23 (1), pp. 85–102, 1984. [citation referenced on p. 65].

Graves, A. "Generating Sequences with Recurrent Neural Networks", ArXiv Preprint ArXiv:1308.0850, 2013. [citation referenced on p. 117].

Grossmann, A. and Morlet, J. "Decomposition of Hardy Functions Into Square Integrable Wavelets of Constant Shape", SIAM Journal on Mathematical Analysis, Vol. 15 (4), pp. 723–736, 1984. [citation referenced on p. 68].

Grossmann, A., Kronland-Martinet, R., and Morlet, J. "Reading and Understanding Continuous Wavelet Transforms", pp. 2–20, 1990. [citation referenced on p. 67, 70].

Guarda-Nardini, L., Manfredini, D., Mion, M., Heir, G., and Marchese-Ragona, R. "Anatomically Based Outcome Predictors of Treatment for Obstructive Sleep Apnea with Intraoral Splint Devices: a Systematic Review of Cephalometric Studies", Journal of Clinical Sleep Medicine, Vol. 11 (11), pp. 1327–1334, 2015. [citation referenced on p. 8].

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. "On Calibration of Modern Neural Networks", pp. 1321–1330, 2017. [citation referenced on p. 142].

Haralick, R. M. "Statistical and Structural Approaches to Texture", Proceedings of the IEEE, Vol. 67 (5), pp. 786–804, 1979. [citation referenced on p. 71, 72, 73].

Project "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". 2009. [citation referenced on p. 99, 100, 105, 106, 108, 112, 113].

Hille, B. "Ionic Channels in Excitable Membranes. Current Problems and Biophysical Approaches", Biophysical Journal, Vol. 22 (2), pp. 283–294, 1978. [citation referenced on p. 12].

Hosselet, J.-J., Norman, R. G., Ayappa, I., and Rapoport, D. M. "Detection of Flow Limitation with a Nasal Cannula/pressure Transducer System", American Journal of Respiratory and Critical Care Medicine, Vol. 157 (5), pp. 1461–1467, 1998. [citation referenced on p. 28].

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. "Snapshot Ensembles: Train 1, Get M for Free", ArXiv Preprint ArXiv:1704.00109, 2017. [citation referenced on p. 121].

Hudgins, B., Parker, P., and Scott, R. N. "A New Strategy for Multifunction Myoelectric Control", IEEE Transactions on Biomedical Engineering, Vol. 40 (1), pp. 82–94, 1993. [citation referenced on p. 51, 82, 83].

Hung, C.-C., Song, E., and Lan, Y. "Image Texture Analysis", Springer, 2019. [[citation referenced on p. 72, 73].

Project "MATLAB Version: 9.13.0 (R2022b)". 2022. [Link: https://www.mathworks.com, citation referenced on p. 74, 87].

Project "Software As a Medical Device (SaMD): Clinical Evaluation". 2017. [citation referenced on p. 131, 190].

Isin, A. and Ozdalili, S. "Cardiac Arrhythmia Detection Using Deep Learning", Procedia Computer Science, 120, pp. 268–275, 2017. [citation referenced on p. 122].

Isono, S., Remmers, J. E., Tanaka, A., Sho, Y., Sato, J., and Nishino, T. "Anatomy of Pharynx in Patients with Obstructive Sleep Apnea and in Normal Subjects", Journal of Applied Physiology, Vol. 82 (4), pp. 1319–1326, 1997. [citation referenced on p. 21].

Isono, S., Tanaka, A., Sho, Y., Konno, A., and Nishino, T. "Advancement of the Mandible Improves Velopharyngeal Airway Patency", Journal of Applied Physiology, Vol. 79 (6), pp. 2132–2138, 1995. [citation referenced on p. 26].

Johnson, K., Schneider, T., Babaian, N., Azage, M., and McIntosh, P. "475P A Novel Clinic to Resolve Variants of Uncertain Significance in Neuromuscular Patients", Neuromuscular Disorders, 43, pp. 104441–312, 2024. [citation referenced on p. 189].

Jordan, A. S., Eckert, D. J., Wellman, A., Trinder, J. A., Malhotra, A., and White, D. P. "Termination of Respiratory Events with and Without Cortical Arousal in Obstructive Sleep Apnea", American Journal of Respiratory and Critical Care Medicine, Vol. 184 (10), pp. 1183–1191, 2011. DOI: 10.1164/rccm.201106-0975oc, citation referenced on p. 22].

Jugé, L., Knapman, F. L., Humburg, P., Burke, P. G., Lowth, A. B., Brown, E., Butler, J. E., Eckert, D. J., Ngiam, J., and Sutherland, K. "The Relationship Between Mandibular Advancement, Tongue Movement, and Treatment Outcome in Obstructive Sleep Apnea", Sleep, Vol. 45 (6), zsac044, 2022. [citation referenced on p. 8, 41].

Julesz, B. "Experiments in the Visual Perception of Texture", Scientific American, Vol. 232 (4), pp. 34–43, 1975. [citation referenced on p. 71].

Jurkat-Rott, K., Lerche, H., and Lehmann-Horn, F. "Skeletal Muscle Channelopathies", Journal of Neurology, 249, pp. 1493–1502, 2002. [citation referenced on p. 11, 13].

Kadambe, S. and Boudreaux-Bartels, G. F. "A Comparison of the Existence Of'cross Terms' in the Wigner Distribution and the Squared Magnitude of the Wavelet Transform and the Short-time Fourier Transform", IEEE Transactions on Signal Processing, Vol. 40 (10), pp. 2498–2517, 1992. [citation referenced on p. 69].

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., and Harrod, C. G. "Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: an American Academy of Sleep Medicine Clinical Practice Guideline", Journal of Clinical Sleep Medicine, Vol. 13 (3), pp. 479–504, 2017. [citation referenced on p. 23, 24].

Kato, J., Isono, S., Tanaka, A., Watanabe, T., Araki, D., Tanzawa, H., and Nishino, T. "Dose-dependent Effects of Mandibular Advancement on Pharyngeal Mechanics and Nocturnal Oxygenation in Patients with Sleep-disordered Breathing", Chest, Vol. 117 (4), pp. 1065–1072, 2000. [citation referenced on p. 26].

Kent, D. T., Rogers, R., and Soose, R. J. "Drug-induced Sedation Endoscopy in the Evaluation of OSA Patients with Incomplete Oral Appliance Therapy Response", Otolaryngology–Head and Neck Surgery, Vol. 153 (2), pp. 302–307, 2015. [citation referenced on p. 29].

Khushaba, R. N., Al-Timemy, A. H., Al-Ani, A., and Al-Jumaily, A. "A Framework of Temporal-spatial Descriptors-based Feature Extraction for Improved Myoelectric Pattern Recognition", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 25 (10), pp. 1821–1831, 2017. [citation referenced on p. 82].

Kimura, J. and Strakowski, J. A. "Electrodiagnosis in Diseases of Nerve and Muscle: Principles and Practice", Oxford university press, 2025. [[citation referenced on p. 15, 16, 17].

Klotz, T., Gizzi, L., Yavuz, U. Ş., and Röhrle, O. "Modelling the Electrical Activity of Skeletal Muscle Tissue Using a Multi-domain Approach", Biomechanics and Modeling in Mechanobiology, Vol. 19 (1), pp. 335–349, 2020. [citation referenced on p. XII, XVI, 34, 155].

Koczwara, K. E., Lake, N. J., DeSimone, A. M., and Lek, M. "Neuromuscular Disorders: Finding the Missing Genetic Diagnoses", Trends in Genetics, Vol. 38 (9), pp. 956–971, 2022. [citation referenced on p. 189].

Kohavi, R. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection", Vol. Vol. 14 (2), pp. 1137–1145, 1995. [citation referenced on p. 100].

Krishnan, S. "Biomedical Signal Analysis for Connected Healthcare", Academic Press, 2021. [[citation referenced on p. 52].

Laine, A. and Fan, J. "Texture Classification by Wavelet Packet Signatures", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15 (11), pp. 1186–1191, 1993. [citation referenced on p. 71].

Lakshminarayanan, B., Pritzel, A., and Blundell, C. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles", Advances in Neural Information Processing Systems, 30, 2017. [citation referenced on p. 122, 141].

LeCun, Y., Kavukcuoglu, K., and Farabet, C. "Convolutional Networks and Applications in Vision", pp. 253–256, 2010. [citation referenced on p. 116].

Lee, H.-K., Kim, H., and Lee, K.-J. "Nasal Pressure Recordings for Automatic Snoring Detection", Medical & Biological Engineering & Computing, 53, pp. 1103–1111, 2015. [citation referenced on p. 28].

Lehmann-Horn, F. and Jurkat-Rott, K. "Voltage-gated Ion Channels and Hereditary Disease", Physiological Reviews, Vol. 79 (4), pp. 1317–1372, 1999. [citation referenced on p. 37, 38].

Levent, E. and Sarıman, N. “Analysis of Obstructive Sleep Apnea Patients with “Sawtooth Sign” on the Flow-volume Curve”, Sleep and Breathing, 15, pp. 357–365, 2011. [citation referenced on p. 28].

Project “Vision in Man and Machine”. 1985. [citation referenced on p. 71].

Lilly, J. M. and Olhede, S. C. “Higher-order Properties of Analytic Wavelets”, IEEE Transactions on Signal Processing, Vol. 57 (1), pp. 146–160, 2008. [citation referenced on p. 65, 66].

Lilly, J. M. and Olhede, S. C. “Generalized Morse Wavelets As a Superfamily of Analytic Wavelets”, IEEE Transactions on Signal Processing, Vol. 60 (11), pp. 6036–6041, 2012. [citation referenced on p. 66].

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. “Focal Loss for Dense Object Detection”, pp. 2980–2988, 2017. [citation referenced on p. 134].

Ling, W.-K. “Nonlinear Digital Filters: Analysis and Applications”, Academic Press, 2010. [[citation referenced on p. 57, 58].

Lu, C. S., Chung, P. C., and Chen, C. F. “Unsupervised Texture Segmentation Via Wavelet Transform”, Pattern Recognition, Vol. 30 (5), pp. 729–742, 1997. [citation referenced on p. 71].

Magalang, U. J. and Grant, B. J. B. “Understanding Stability of Obstructive Sleep Apnea Endotypes: a Step Forward”, Sleep, Vol. 45 (9), zsac174, 2022. DOI: 10.1093/sleep/zsac174, citation referenced on p. 22].

Maggi, L., Ravaglia, S., Farinato, A., Brugnoni, R., Altamura, C., Imbrici, P., Camerino, D. C., Padovani, A., Mantegazza, R., and Bernasconi, P. “Coexistence of CLCN1 and SCN4A Mutations in One Family Suffering From Myotonia”, Neurogenetics, 18, pp. 219–225, 2017. [citation referenced on p. 189].

Malhotra, A. and White, D. P. “Obstructive Sleep Apnoea”, The Lancet, Vol. 360 (9328), pp. 237–245, 2002. [citation referenced on p. 21].

Mallat, S. G. “A Theory for Multiresolution Signal Decomposition: the Wavelet Representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11 (7), pp. 674–693, 1989. [citation referenced on p. 71].

Manetta, I. P., Ettlin, D., Sanz, P. M., Rocha, I., and Cruz, M. M. “Mandibular Advancement Devices in Obstructive Sleep Apnea: an Updated Review”, Sleep Science, Vol. 15 (S 02), pp. 398–405, 2022. [citation referenced on p. 25].

Mann, D. L., Terrill, P. I., Azarbarzin, A., Mariani, S., Franciosini, A., Camassa, A., Georgeson, T., Marques, M., Taranto-Montemurro, L., and Messineo, L. “Quantifying the Magnitude of Pharyngeal Obstruction During Sleep Using Airflow Shape”, European Respiratory Journal, Vol. 54 (1), 2019. [citation referenced on p. 27, 44].

Markhorst, J. M., Stunnenberg, B. C., Ginjaar, I. B., Drost, G., Erasmus, C. E., and Sie, L. T. “Clinical Experience with Long-term Acetazolamide Treatment in Children with Nondystrophic Myotonias: a Three-case Report”, Pediatric Neurology, Vol. 51 (4), pp. 537–541, 2014. [citation referenced on p. 3].

Martinez-Ríos, E. A., Bustamante-Bello, R., Navarro-Tuch, S., and Perez-Meana, H. “Applications of the Generalized Morse Wavelets: a Review”, IEEE Access, 11, pp. 667–688, 2022. [citation referenced on p. 66, 67].

Materka, A. and Strzelecki, M. "Texture Analysis Methods–a Review", Technical University of Lodz, Institute of Electronics, COST B11 Report, Brussels, Vol. 10 (1.97), p. 4968, 1998. [citation referenced on p. 71].

Matthews, E., Fialho, D., Tan, S., Venance, S., Cannon, S., Sternberg, D., Fontaine, B., Amato, A., Barohn, R., and Griggs, R. "The Non-dystrophic Myotonias: Molecular Pathogenesis, Diagnosis and Treatment", Brain, Vol. 133 (1), pp. 9–22, 2010. [citation referenced on p. 2, 3, 37, 38, 39].

Project "The Pursuit of Oral Appliance Response Predictors". 2022. [citation referenced on p. 7].

McManis, P. G., Lambert, E. H., and Daube, J. R. "The Exercise Test in Periodic Paralysis", Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, Vol. 9 (8), pp. 704–710, 1986. [citation referenced on p. 37].

Midrio, M. "The Denervated Muscle: Facts and Hypotheses. A Historical Review", European Journal of Applied Physiology, 98, pp. 1–21, 2006. [citation referenced on p. 16].

Mitrović, N., George Jr, A., Lerche, H., Wagner, S., Fahlke, C., and Lehmann-Horn, F. "Different Effects on Gating of Three Myotonia-causing Mutations in the Inactivation Gate of the Human Muscle Sodium Channel." The Journal of Physiology, Vol. 487 (1), pp. 107–114, 1995. [citation referenced on p. 12, 18, 37].

Mohammadieh, A., Tong, B., De Chazal, P., and Cistulli, P. A. "Innovations in Mandibular Advancement Splint Therapy for Obstructive Sleep Apnoea", Frontiers in Sleep, 2, p. 1144327, 2023. [citation referenced on p. 25].

Morales, F. and Pusch, M. "An Up-to-date Overview of the Complexity of Genotype-phenotype Relationships in Myotonic Channelopathies", Frontiers in Neurology, 10, p. 1404, 2020. [citation referenced on p. 2, 3].

Moyer, C. A., Sonnad, S. S., Garetz, S. L., Helman, J. I., and Chervin, R. D. "Quality of Life in Obstructive Sleep Apnea: a Systematic Review of the Literature", Sleep Medicine, Vol. 2 (6), pp. 477–491, 2001. [citation referenced on p. 19].

Nam, S., Sohn, M. K., Kim, H. A., Kong, H.-J., and Jung, I.-Y. "Development of Artificial Intelligence to Support Needle Electromyography Diagnostic Analysis", Healthcare Informatics Research, Vol. 25 (2), pp. 131–138, 2019. [citation referenced on p. 5].

Nandedkar, S. D. and Barkhaus, P. E. "Quantitative EMG Analysis", Springer, pp. 165–199, 2013. [citation referenced on p. 3].

Negi, P. C., Pandey, S., Sharma, S., and Sharma, N. "Classification of Gait Abnormalities Using Transfer Learning with EMG Scalogram Features", pp. 407–415, 2023. [citation referenced on p. 4].

Ng, A. T., Qian, J., and Cistulli, P. A. "Oropharyngeal Collapse Predicts Treatment Response with Oral Appliance Therapy in Obstructive Sleep Apnea", Sleep, Vol. 29 (5), pp. 666–671, 2006. [citation referenced on p. 29].

Nilsson, J., Panizza, M., and Hallett, M. "Principles of Digital Sampling of a Physiologic Signal", Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, Vol. 89 (5), pp. 349–358, 1993. [citation referenced on p. 15, 55, 60].

Nodera, H., Osaki, Y., Yamazaki, H., Mori, A., Izumi, Y., and Kaji, R. "Deep Learning for Waveform Identification of Resting Needle Electromyography Signals", Clinical Neurophysiology, Vol. 130 (5), pp. 617–623, 2019. [citation referenced on p. 5].

Oh, S. J. "Clinical Electromyography: Nerve Conduction Studies", Lippincott Williams & Wilkins, 2003. [[citation referenced on p. 17].

Olhede, S. C. and Walden, A. T. "Generalized Morse Wavelets", IEEE Transactions on Signal Processing, Vol. 50 (11), pp. 2661–2670, 2002. [citation referenced on p. 66].

Oppenheim, A. V. "Discrete-time Signal Processing", Pearson Education India, 1999. [[citation referenced on p. 54, 184].

"Oronasal Vs Nasal Masks: The Impact of Mask Type on CPAP Requirement, Residual AHI, and Adherence", CHEST, 2023. DOI: 10.1016/j.chest.2023.06.xxx, citation referenced on p. 25].

Pappone, P. A. "Voltage-clamp Experiments in Normal and Denervated Mammalian Skeletal Muscle Fibres." The Journal of Physiology, Vol. 306 (1), pp. 377–410, 1980. [citation referenced on p. 16].

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. "Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition", ArXiv Preprint ArXiv:1904.08779, 2019. [citation referenced on p. 125].

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. "Automatic Differentiation in PyTorch", 2017. [citation referenced on p. 99].

Patil, S. P., Ayappa, I., Caples, S. M., Kimoff, R. J., Patel, S. R., and Harrod, C. G. "Treatment of Adult Obstructive Sleep Apnea with Positive Airway Pressure: An American Academy of Sleep Medicine Clinical Practice Guideline", Journal of Clinical Sleep Medicine, Vol. 15 (2), pp. 335–343, 2019. DOI: 10.5664/jcsm.7640, citation referenced on p. 25].

Pattnaik, S., Rao, B. N., Rout, N. K., and Sabut, S. K. "Transfer Learning Based Epileptic Seizure Classification Using Scalogram Images of EEG Signals", Multimedia Tools and Applications, Vol. 83 (36), pp. 84179–84193, 2024. [citation referenced on p. 4].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, 12, pp. 2825–2830, 2011. [citation referenced on p. 99].

Percival, D. B. and Walden, A. T. "Spectral Analysis for Physical Applications", cambridge university press, 1993. [[citation referenced on p. 61].

Pham, S. and Puckett, Y. "Physiology, Skeletal Muscle Contraction", 2020. [citation referenced on p. 11, 12].

Phillips, C. L., Grunstein, R. R., Darendeliler, M. A., Mihailidou, A. S., Srinivasan, V. K., Yee, B. J., Marks, G. B., and Cistulli, P. A. "Health Outcomes of Continuous Positive Airway Pressure Versus Oral Appliance Treatment for Obstructive Sleep Apnea: a Randomized Controlled Trial", American Journal of Respiratory and Critical Care Medicine, Vol. 187 (8), pp. 879–887, 2013. [citation referenced on p. 27].

Phinyomark, A., Phukpattaranont, P., and Limsakul, C. "Fractal Analysis Features for Weak and Single-channel Upper-limb EMG Signals", Expert Systems with Applications, Vol. 39 (12), pp. 11156–11163, 2012. [citation referenced on p. 83].

Platt, J. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", Advances in Large Margin Classifiers, Vol. 10 (3), pp. 61–74, 1999. [citation referenced on p. 179].

Pond, A., Marcante, A., Zanato, R., Martino, L., Stramare, R., Vindigni, V., Zampieri, S., Hofer, C., Kern, H., and Masiero, S. "History, Mechanisms and Clinical Value of Fibrillation Analyses in Muscle Denervation and Reinnervation by Single Fiber Electromyography and Dynamic Echomyography", European Journal of Translational Myology, Vol. 24 (1), p. 3297, 2014. [citation referenced on p. 17].

Preston, D. C. and Shapiro, B. E. "Needle Electromyography: Fundamentals, Normal and Abnormal Patterns", Neurologic Clinics, Vol. 20 (2), pp. 361–396, 2002. [citation referenced on p. 18].

Preston, D. C. and Shapiro, B. E. "Electromyography and Neuromuscular Disorders E-book: Clinical-electrophysiologic Correlations (Expert Consult-Online)", Elsevier Health Sciences, 2012. [[citation referenced on p. 17].

Proakis, J. G. "Digital Signal Processing: Principles Algorithms and Applications", Pearson Education India, 2001. [[citation referenced on p. 57, 58, 184].

Python Core Team, "Python: A dynamic, open source programming language", 2019 Python version 3.7.[Link: https://www.python.org/, citation referenced on p. 95, 99, 174].

Rioul, O. and Vetterli, M. "Wavelets and Signal Processing", IEEE Signal Processing Magazine, Vol. 8 (4), pp. 14–38, 1991. [citation referenced on p. 61, 62, 63, 64, 67, 68, 69, 70].

Project "Genetic Testing & Neuromuscular Disorders". 2022. [citation referenced on p. 189].

Rosenfeld, A. "Digital Picture Processing", Academic press, 1976. [[citation referenced on p. 71].

Rotenberg, B. W., Murariu, D., and Pang, K. P. "Trends in CPAP Adherence Over Twenty Years of Data Collection: a Flattened Curve", Journal of Otolaryngology—Head & Neck Surgery, Vol. 45 (1), p. 43, 2016. DOI: 10.1186/s40463-016-0156-0, citation referenced on p. 25].

Rotty, M.-C., Suehs, C. M., and Mallet, J.-P. "Mask Side-effects in Long-term CPAP-patients Impact Adherence and Sleepiness: the InterfaceVent Real-life Study", Respiratory Research, 22, p. 17, 2021. DOI: 10.1186/s12931-021-01618-x, citation referenced on p. 25].

Rubin, D. I. "Normal and Abnormal Spontaneous Activity", Handbook of Clinical Neurology, 160, pp. 257–279, 2019. [citation referenced on p. 16].

Ryan, C., Love, L., Peat, D., Fleetham, J., and Lowe, A. "Mandibular Advancement Oral Appliance Therapy for Obstructive Sleep Apnoea: Effect on Awake Calibre of the Velopharynx", Thorax, Vol. 54 (11), pp. 972–977, 1999. [citation referenced on p. 25].

Salem, M., Taheri, S., and Yuan, J.-.-S. "ECG Arrhythmia Classification Using Transfer Learning From 2-Dimensional Deep CNN Features", pp. 1–4, 2018. [citation referenced on p. 4].

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. "Grad-cam: Visual Explanations From Deep Networks Via Gradient-based Localization", pp. 618–626, 2017. [citation referenced on p. 147].

Shannon, C. E. "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol. 27 (3), pp. 379–423, 1948. [citation referenced on p. 48].

Shorten, P. R., O'Callaghan, P., Davidson, J. B., and Soboleva, T. K. "A Mathematical Model of Fatigue in Skeletal Muscle Force Contraction", Journal of Muscle Research and Cell Motility, 28, pp. 293–313, 2007. [citation referenced on p. 34].

Slepian, D. "Some Comments on Fourier Analysis, Uncertainty and Modeling", SIAM Review, Vol. 25 (3), pp. 379–393, 1983. [citation referenced on p. 61].

Smith, S. W. "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Pub, 1997. [citation referenced on p. 49, 53, 54, 55, 56, 57, 60, 70, 71].

Sparber, P., Sharova, M., Filatova, A., Shchagina, O., Ivanova, E., Dadali, E., and Skoblov, M. "Recessive Myotonia Congenita Caused by a Homozygous Splice Site Variant in CLCN1 Gene: a Case Report", BMC Medical Genetics, 21, pp. 1–5, 2020. [citation referenced on p. 189].

Stålberg, E., Dijk, H., Falck, B., Kimura, J., Neuwirth, C., Pitt, M., Podnar, S., Rubin, D. I., Rutkove, S., and Sanders, D. B. "Standards for Quantification of EMG and Neurography", Clinical Neurophysiology, Vol. 130 (9), pp. 1688–1729, 2019. [citation referenced on p. 15, 16].

STARD-AI Steering Group, "The STARD-AI Reporting Guideline for Diagnostic Accuracy Studies Using Artificial Intelligence", Nature Medicine, 2025. DOI: 10.1038/s41591-025-03953-8, citation referenced on p. 190].

Statland, J. M., Bundy, B. N., Wang, Y., Rayan, D. R., Trivedi, J. R., Sansone, V. A., Salajegheh, M. K., Venance, S. L., Ciafaloni, E., and Matthews, E. "Mexiletine for Symptoms and Signs of Myotonia in Nondystrophic Myotonia: a Randomized Controlled Trial", Jama, Vol. 308 (13), pp. 1357–1365, 2012. [citation referenced on p. 14].

Subasi, A. "Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques: A MATLAB Based Approach", Academic Press, 2019. [[citation referenced on p. 16].

Suetterlin, K., Matthews, E., Sud, R., McCall, S., Fialho, D., Burge, J., Jayaseelan, D., Haworth, A., Sweeney, M. G., and Kullmann, D. M. "Translating Genetic and Functional Data Into Clinical Practice: a Series of 223 Families with Myotonia", Brain, Vol. 145 (2), pp. 607–620, 2022. [citation referenced on p. 189].

Sun, L.-C., Lee, C.-C., Ke, H.-Y., Wei, C.-Y., Lin, K.-F., Lin, S.-S., Hsiu, H., and Chen, P.-N. "Deep Learning for the Classification of Atrial Fibrillation Using Wavelet Transform-based Visual Images", BMC Medical Informatics and Decision Making, Vol. 22 (Suppl 5), p. 349, 2025. [citation referenced on p. 4].

Sun, X., Liu, P., He, Z., Han, Y., and Su, B. "Automatic Classification of Electrocardiogram Signals Based on Transfer Learning and Continuous Wavelet Transform", Ecological Informatics, 69, p. 101628, 2022. [citation referenced on p. 4].

Sutherland, K. and Cistulli, P. "Mandibular Advancement Splints for the Treatment of Sleep Apnoea Syndrome", Swiss Medical Weekly, Vol. 141 (3940), w13276–w13276, 2011. [citation referenced on p. 7].

Sutherland, K., Vanderveken, O. M., Tsuda, H., Marklund, M., Gagnadoux, F., Kushida, C. A., and Cistulli, P. A. "Oral Appliance Treatment for Obstructive Sleep Apnea: an Update", Journal of Clinical Sleep Medicine, Vol. 10 (2), pp. 215–227, 2014. [citation referenced on p. 7, 8, 19, 26, 27].

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. "Rethinking the Inception Architecture for Computer Vision", pp. 2818–2826, 2016. [citation referenced on p. 134].

Tang, C.-Y. and Chen, T.-Y. "Physiology and Pathophysiology of CLC-1: Mechanisms of a Chloride Channel Disease, Myotonia", BioMed Research International, Vol. 2011 (1), p. 685328, 2011. [citation referenced on p. 13].

Tankisi, H., Burke, D., Cui, L., De Carvalho, M., Kuwabara, S., Nandedkar, S. D., Rutkove, S., Stålberg, E., Putten, M. J., and Fuglsang-Frederiksen, A. "Standards of Instrumentation of EMG", Clinical Neurophysiology, Vol. 131 (1), pp. 243–258, 2020. [citation referenced on p. 15].

Thesleff, S. and Ward, M. "Studies on the Mechanism of Fibrillation Potentials in Denervated Muscle." The Journal of Physiology, Vol. 244 (2), pp. 313–323, 1975. [citation referenced on p. 16].

Thornton, R. C. and Michell, A. W. "Techniques and Applications of EMG: Measuring Motor Units From Structure to Function", Journal of Neurology, 259, pp. 585–594, 2012. [citation referenced on p. 17].

Too, J., Abdullah, A. R., and Saad, N. M. "Classification of Hand Movements Based on Discrete Wavelet Transform and Enhanced Feature Extraction", International Journal of Advanced Computer Science and Applications, Vol. 10 (6), 2019. [citation referenced on p. 81, 82].

Trip, J., Drost, G., Ginjaar, H., Nieman, F., Van Der Kooi, A., Visser, M., Engelen, B., and Faber, C. "Redefining the Clinical Phenotypes of Non-dystrophic Myotonic Syndromes", Journal of Neurology, Neurosurgery & Psychiatry, Vol. 80 (6), pp. 647–652, 2009. [citation referenced on p. 14].

Trip, J., Drost, G., Verbove, D. J., Van Der Kooi, A. J., Kuks, J., Notermans, N. C., Verschuuren, J. J., De Visser, M., Van Engelen, B. G., and Faber, C. G. "In Tandem Analysis of CLCN1 and SCN4A Greatly Enhances Mutation Detection in Families with Non-dystrophic Myotonia", European Journal of Human Genetics, Vol. 16 (8), pp. 921–929, 2008. [citation referenced on p. 2, 3, 6, 12, 32, 37, 38, 39].

Project "Statistical Guidance on Reporting Results From Studies Evaluating Diagnostic Tests: Guidance for Industry and FDA Staff". 2007. [citation referenced on p. 131].

Project "De Novo Summary (DEN180001): IDx-DR". 2018. [citation referenced on p. 131].

Vacchiano, V., Brugnoni, R., Campanale, C., Imbrici, P., Dinoi, G., Canioni, E., Laghetti, P., Saltarella, I., Altamura, C., and Maggi, L. "Coexistence of SCN4A and CLCN1 Mutations in a Family with Atypical Myotonic Features: A Clinical and Functional Study", Experimental Neurology, 362, p. 114342, 2023. [citation referenced on p. 189].

Vapnik, V. and Vapnik, V. "Statistical Learning Theory Wiley", New York, Vol. 1 (624), p. 2, 1998. [citation referenced on p. 122].

Vapnik, V. N. and Chervonenkis, A. Y. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities", Springer, pp. 11–30, 2015. [citation referenced on p. 123].

Vaswani, A. "Attention is all You Need", Advances in Neural Information Processing Systems, 2017. [citation referenced on p. 117, 118].

Vedantham, V. and Cannon, S. C. "Rapid and Slow Voltage-dependent Conformational Changes in Segment IVS6 of Voltage-gated Na+ Channels", Biophysical Journal, Vol. 78 (6), pp. 2943–2958, 2000. [citation referenced on p. 4].

Vena, D., Azarbarzin, A., Marques, M., Op de Beeck, S., Vanderveken, O. M., Edwards, B. A., Calianese, N., Hess, L. B., Radmand, R., and Hamilton, G. S. "Predicting Sleep

Apnea Responses to Oral Appliance Therapy Using Polysomnographic Airflow", Sleep, Vol. 43 (7), zsaa004, 2020. [citation referenced on p. 9].

Vereb, N., Montagnese, F., Gläser, D., and Schoser, B. "Non-dystrophic Myotonias: Clinical and Mutation Spectrum of 70 German Patients", Journal of Neurology, 268, pp. 1708–1720, 2021. [citation referenced on p. 2].

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. 1. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", Nature Methods, 2020. DOI: https://doi.org/10.1038/s41592-019-0686-2, citation referenced on p. 81].

Wagner, S., Deymeer, F., Kürz, L. L., Benz, S., Schleithoff, L., Lehmann-Horn, F., Serdaroğlu, P., Özdemir, C., and Rüdel, R. "The Dominant Chloride Channel Mutant G200R Causing Fluctuating Myotonia: Clinical Findings, Electrophysiology, and Channel Pathology", Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, Vol. 21 (9), pp. 1122–1128, 1998. [citation referenced on p. 37].

Weaver, T. E. and Grunstein, R. "Adherence to Continuous Positive Airway Pressure Therapy", Proceedings of the American Thoracic Society, Vol. 5 (2), pp. 173–178, 2008. DOI: 10.1513/pats.200708-119mg, citation referenced on p. 25].

Welch, P. "The Use of Fast Fourier Transform for the Estimation of Power Spectra: a Method Based on Time Averaging Over Short, Modified Periodograms", IEEE Transactions on Audio and Electroacoustics, Vol. 15 (2), pp. 70–73, 1967. [citation referenced on p. 58, 59].

Weszka, J. S., Dyer, C. R., and Rosenfeld, A. "A Comparative Study of Texture Measures for Terrain Classification", IEEE Transactions on Systems, Man, and Cybernetics, pp. 269–285, 1976. [citation referenced on p. 71].

Willmott, A. D., White, C., and Dukelow, S. P. "Fibrillation Potential Onset in Peripheral Nerve Injury", Muscle & Nerve, Vol. 46 (3), pp. 332–340, 2012. [citation referenced on p. 16].

Woyczynski, W. A. and Woyczyânski, W. "A First Course in Statistics for Signal Analysis", Springer, 2011. [[citation referenced on p. 66].

Wu, F.-F., Ryan, A., Devaney, J., Warnstedt, M., Korade-Mirnics, Z., Poser, B., Escriva, M. J., Pegoraro, E., Yee, A. S., and Felice, K. J. "Novel CLCN1 Mutations with Unique Clinical and Electrophysiological Consequences", Brain, Vol. 125 (11), pp. 2392–2407, 2002. [citation referenced on p. 189].

Yeghiazarians, Y., Jneid, H., Tietjens, J. R., Redline, S., Brown, D. L., El-Sherif, N., Mehra, R., Bozkurt, B., Ndumele, C. E., and Somers, V. K. "Obstructive Sleep Apnea and Cardiovascular Disease: A Scientific Statement From the American Heart Association", Circulation, Vol. 144 (3), e56–e67, 2021. DOI: 10.1161/cir.0000000000000988, citation referenced on p. 22].

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. "How Transferable Are Features in Deep Neural Networks?", Advances in Neural Information Processing Systems, 27, 2014. [citation referenced on p. 119].

Zadrozny, B. and Elkan, C. "Transforming Classifier Scores Into Accurate Multiclass Probability Estimates", pp. 694–699, 2002. [citation referenced on p. 181].

Zeiler, M. D. and Fergus, R. “Visualizing and Understanding Convolutional Networks”, pp. 818–833, 2014. [citation referenced on p. 120].

Zeng, B., Ng, A. T., Darendeliler, M. A., Petocz, P., and Cistulli, P. A. “Use of Flow–volume Curves to Predict Oral Appliance Treatment Outcome in Obstructive Sleep Apnea”, American Journal of Respiratory and Critical Care Medicine, Vol. 175 (7), pp. 726–730, 2007. [citation referenced on p. 9].

Zhang, H. “Mixup: Beyond Empirical Risk Minimization”, ArXiv Preprint ArXiv:1710.09412, 2017. [citation referenced on p. 123, 124].

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. “Random Erasing Data Augmentation”, Vol. Vol. 34 (07), pp. 13001–13008, 2020. [citation referenced on p. 124].