machines

MDPI

Article

Uncertainty Quantification for Full-Flight Data Based Engine Fault Detection with Neural Networks

Matthias Weiss [1,*], Stephan Staudacher [1], Jürgen Mathes [2], Duilio Becchio [2] and Christian Keller [3]

[1] Institute of Aircraft Propulsion Systems, University of Stuttgart, 70569 Stuttgart, Germany
[2] MTU Aero Engines AG, 80995 München, Germany
[3] MTU Maintenance Hannover GmbH, 30855 Langenhagen, Germany
[*] Correspondence: matthias.weiss@ila.uni-stuttgart.de; Tel.: +49-711-685-69383

Abstract: Current state-of-the-art engine condition monitoring is based on a minimum of one steady-state data point per flight. Due to the scarcity of available data points, there are difficulties distinguishing between random scatter and an underlying fault introducing a detection latency of several flights. Today's increased availability of data acquisition hardware in modern aircraft provides continuously sampled in-flight measurements, so-called full-flight data. These full-flight data give access to sufficient data points to detect faults within a single flight, significantly improving the availability and safety of aircraft. Artificial neural networks are considered well suited for the timely analysis of an extensive amount of incoming data. This article proposes uncertainty quantification for artificial neural networks, leading to more reliable and robust fault detection. An existing approach for approximating the aleatoric uncertainty was extended by an Out-of-Distribution Detection in order to take the epistemic uncertainty into account. The method was statistically evaluated, and a grid search was performed to evaluate optimal parameter combinations maximizing the true positive detection rates. All test cases were derived based on in-flight measurements of a commercially operated regional jet. Especially when requiring low false positive detection rates, the true positive detections could be improved 2.8 times while improving response times by approximately 6.9 compared to methods only accounting for the aleatoric uncertainty.

Keywords: aircraft engine; gas turbine; fault detection; engine health monitoring; engine condition monitoring; full-flight data; artificial neural networks; uncertainty quantification

Citation: Weiss, M.; Staudacher, S.; Mathes, J.; Becchio, D.; Keller, C. Uncertainty Quantification for Full-Flight Data Based Engine Fault Detection with Neural Networks. *Machines* **2022**, *10*, 846. https://doi.org/10.3390/machines10100846

Academic Editors: Konstantinos Kyprianidis and Valentina Zaccaria

Received: 22 August 2022
Accepted: 20 September 2022
Published: 23 September 2022

1. Introduction

Engine condition monitoring is considered a key technology for lowering maintenance, repair and overhaul expenses while improving the safety and availability of aircraft [1]. Estimating the current health state of the aircraft engine gained from engine condition monitoring systems by analyzing in-flight measurements provides the foundation for effective maintenance planning. Besides tracking and trending long-term deterioration, engine condition monitoring applications detect, isolate and identify single faults [2].

Current state-of-the-art engine condition monitoring systems i.e., Refs. [3–6] are based on analyzing a minimum of one steady-state snapshot per flight. The sparsity of available data negatively impacts fault detection as there are difficulties distinguishing between random scatter and an underlying fault. Depending on the fault type and severity, it can take several flights until fault detection [5,7,8]. The resulting latency in fault detection increases the risk of secondary damage. Recently, with the increased adoption of non-mandatory data acquisition equipment, continuously sampled datasets are available covering whole flights. These continuously sampled datasets are also referred to as full-flight data. Full-flight data provide sufficient data points to detect engine faults based on a statistically relevant sample size within a single flight, enabling faster response times. Despite the advantages of full-flight data, analyzing the corresponding datasets heavily increases the amount of data to be processed [9]. For timely analysis of the increased number of incoming data, novel algorithms are required.

One approach to improve analysis performance is to reduce data by focusing on representative data points within steady-state operating regimes. The utilization of a linearized state-space model for fault detection in combination with a Kalman Filter for the isolation and identification of the fault is described in [10,11]. An alternative approach combining a steady-state data filter with a thermodynamic engine model and a Once-Class Support Vector Machine for fault detection is proposed in [12]. These methods work well for flights with extended cruise segments where many steady-state operating regimes can be identified. For short-haul flights without extended cruise segments, on the other hand, the total number of identified steady-state data points might be insufficient for fault detection. The results of the steady-state data filter presented in [12] applied to two example flights are visualized in Figure 1 to emphasize this issue. In order to perform fault detection for short-haul flights, alternative approaches are required that analyze the entire flight, including transient operating regimes.

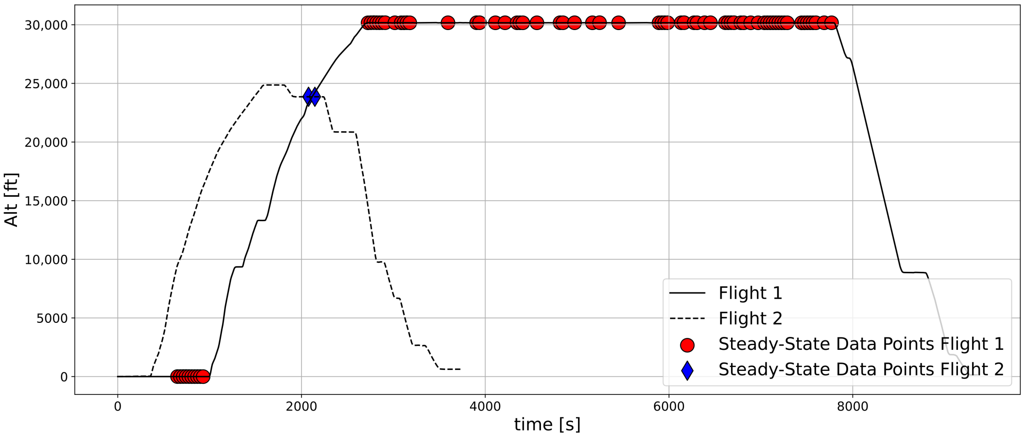

Figure 1. Identified steady-state data points for two example flights.

According to [13] information redundancy is required for fault detection and diagnosis. In current engine condition monitoring applications, this redundancy is typically established by utilizing thermodynamic engine models for computing reference values representing the nominal performance of the aircraft engine. Fault detection performs a comparison between these reference values and in-flight measurements. Significant deviations between the measurements and model predictions indicate an underlying fault. In general, fast execution times are required to analyze the large number of data points provided by full-flight data. Thermodynamic engine models are generally slow since the solution is determined iteratively. On the other hand, state-of-the-art machine learning approaches are well suited for analyzing full-flight data providing fast execution times omitting the slow iterative computation of thermodynamic engine models. Depending on the configuration of the data acquisition, full-flight data often include discrete features resembling the position of valves, e.g., for anti-icing and customer bleed extraction. Building a physically sound thermodynamic engine model without profound system information is difficult as a meaningful relationship between discrete parameter setting and mass flow extraction has to be derived. On the other hand, data-driven models can infer the effect of such discrete parameters. The sometimes limited system information, in combination with the requirement for timely data analysis, makes data-driven model building a good alternative for processing full-flight data.

Different data-driven methods such as artificial neural networks [14–19], Generalized Additive Models [17,19,20] or Support Vector Regression [21] have already been successfully applied to model the performance of gas turbines. However, one major drawback of data-driven approaches is their black-box characteristic making it difficult to substantiate the results. Especially the widespread utilization of artificial neural networks also covering safety-critical applications, e.g., self-driving cars [22], or medical diagnosis [23] lead to increased research in uncertainty quantification, improving the reliability and robustness of their results.

In general, two types of uncertainty are differentiated in model building: aleatoric uncertainty and epistemic uncertainty [24]. Aleatoric uncertainty defines the inherently probabilistic variability of a dataset caused by measurement uncertainty. On the other hand, epistemic uncertainty defines the uncertainty caused by the insufficient coverage of the relevant value range by the available data. For example, when using artificial neural networks for approximating the input-output characteristic of a technical system, they basically define a high-dimensional curve fit. However, the output of the artificial neural network is essentially only trustworthy in operating regimes for which sufficient data have been available for training. Otherwise, the extrapolation error becomes dominant [25,26]. While the epistemic uncertainty can be minimized by taking additional data points of different operating regimes into account, the aleatoric uncertainty is more or less fixed. Dedicated algorithms handle the approximation of the aleatoric and epistemic uncertainty. The epistemic uncertainty can be approximated, for example utilizing Ensemble Models [27], Out-of-Distribution Detection [28], Dropout [29] or Bayesian Neural Networks [30]. The aleatoric uncertainty can be evaluated by approximating the probability density functions of individual measurements with artificial neural networks [31]. Despite an existing concept for approximating the aleatoric uncertainty for full-flight engine data [32], there is no method taking both the aleatoric and epistemic uncertainty into account.

In the following, artificial neural networks are chosen for approximating the performance of aircraft engines. Correctly assessing the temporal correlations in full-flight data is a prerequisite for approximating the engine performance [33] and is more difficult to achieve with other data-driven modeling methods. Amongst artificial neural networks, there are specific architectures to process time series, such as Long-Sort Term Memory (LSTM) [34], Gated Recurrent Units (GRU) [35], or Dilated Convolutional Neural Networks [36]. Apart from the proven capability of the above listed artificial neural networks to model the steady-state and transient performance of gas turbines, there is additionally existing research in uncertainty quantification for neural networks. One existing method for approximating the aleatoric uncertainty in [32] is extended by an Out-of-Distribution Detection for additionally taking the epistemic uncertainty into account. The proposed approach is then tested utilizing full-flight data of a commercially operated regional jet. A comprehensive investigation of the detection rates underlying different fault cases is provided. With the results obtained, it can be shown that the additional uncertainty quantification leads to higher detection rates with faster response times.

2. Materials and Methods

2.1. Artificial Neural Networks with Uncertainty Quantification

Since the approximation of the aleatoric uncertainty according to [32] has already been successfully applied to in-flight measurements, it is used as starting point for further improvement. The approximation of the aleatoric uncertainty introduces additional model complexity to the neural network by requiring an increased number of output nodes. Therefore, a complementary method for estimating epistemic uncertainty was chosen, leaving the artificial neural network unchanged. Of the methods listed in the previous section, only the Out-of-Distribution Detection meets these requirements.

2.1.1. Approximating the Aleatoric Uncertainty

For modeling the aleatoric uncertainty, the training data are assumed to be sampled from a given probability density function $p(y|\Theta)$ with parameters Θ. The parameters Θ of the probability density function are then estimated by the neural network based on input parameters x.

For example, utilizing a Gaussian probability density function in Equation (1) for approximating the probability distribution of the measurements y requires the mean μ and the standard deviation σ to be approximated by the artificial neural network. The parameters are estimated by defining the corresponding output nodes of an artificial neural network. An example of the resulting architecture of the artificial neural network underlying a Gaussian probability function is visualized in Figure 2.

$$p(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \tag{1}$$

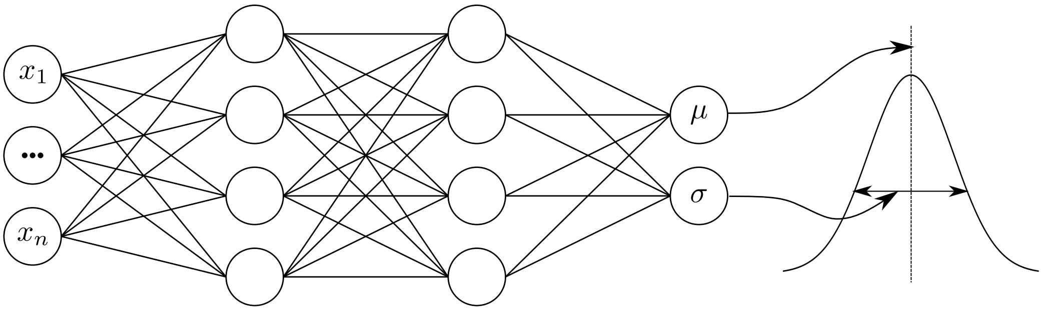

Figure 2. Architecture for approximating a univariate Gaussian probability density function with artificial neural networks.

An optimization defines the weights and biases of the neural network nodes, maximizing the likelihood of observing the data underlying the chosen probability density function $p(y|\Theta)$. Concerning the objective function of the optimization, maximizing the likelihood is equal to minimizing the negative log-likelihood $\mathcal{NLL}$. For a flight of length l, the corresponding negative log-likelihood is defined by Equation (2) [31].

$$\mathcal{NLL} = -log\left(\prod_{j=0}^{l} p(y_j|\Theta_j)\right) = -\sum_{j=0}^{l} log\, p(y_j|\Theta_j) \tag{2}$$

Especially for long-haul flights with extended cruise segments, there are more data points for cruise than other flight segments. This imbalance in data can bias the neural network towards approximating the cruise with high accuracy while neglecting the remaining flight segments. In order to ensure that all flight phases are represented with similar accuracy, the negative log-likelihood $\mathcal{NLL}$ is first computed for each flight phase separately. The optimization is then based on the average negative log-likelihood $\mathcal{NLL}$.

In general, engine condition monitoring requires multivariate datasets to be estimated for which the approach presented above can easily be extended. However, the approximation of multivariate datasets increases the total number of parameters Θ to be estimated as additional cross-correlations between variables must be considered. In the present work, the in-flight measurements are approximated assuming multivariate Gaussian probability density functions, which results in

$$p(\vec{y}|\vec{\mu},\mathbf{\Sigma}) = \frac{1}{|\mathbf{\Sigma}|\sqrt{(2\pi)^n}}\, exp\left(-\frac{1}{2}(\vec{y}-\vec{\mu})^T\mathbf{\Sigma}(\vec{y}-\vec{\mu})\right) \tag{3}$$

$$\vec{\mu} = [\mu_1, \cdots, \mu_i, \cdots, \mu_n]^T \tag{4}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix} \tag{5}$$

Even though the correlation matrix $\mathbf{\Sigma}$ is symmetric, i.e., $\Sigma_{i,j} = \Sigma_{j,i}$, the additional cross-correlations $\Sigma_{i,j}$ increase the complexity of the artificial neural network as additional output nodes have to be provided for their estimation. To reduce the total number of parameters to be estimated, the in-flight measurements are considered to be sampled independently, leading to uncorrelated measurement noise and, therefore, negligible cross-correlations $\Sigma_{i,j}$. This simplification collapses the correlation-matrix $\mathbf{\Sigma}$ into a diagonal matrix $\mathbf{\Sigma} = diag(\Sigma_{1,1}, \cdots, \Sigma_{n,n})$.

Correctly assessing the transient performance of aircraft engines requires the previous data points to be considered [33] resulting in an auto-correlation. In order to account for this temporal correlation, a temporal feature extraction utilizing dilated convolutional neural networks [36] is used as a preprocessing step. The resulting architecture of the neural network for approximating the in-flight measurements of aircraft engines is visualized in Figure 3. Input to the artificial neural network is a multivariate time series consisting of continuous and discrete parameters defining the environmental conditions, power settings, and controller settings. In the next step, global feature extraction is conducted by nonlinearly extracting and compressing the temporal information of the provided time series. Finally, the extracted features are processed by individual feed-forward neural networks approximating the measurements' mean μ and standard deviation σ. The feature extraction and the neural network for estimating the probability density function are trained simultaneously. A similar approach for estimating the aleatoric uncertainty applied to full-flight data is discussed in [32].

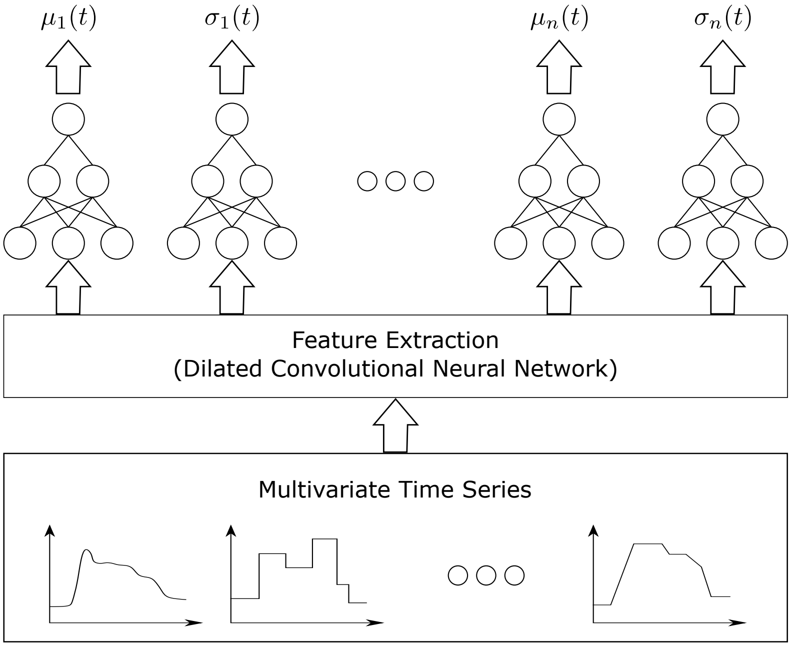

Figure 3. Architecture of the neural network used for approximating the in-flight measurements of aircraft engines.

2.1.2. Approximating the Epistemic Uncertainty

The epistemic uncertainty of neural networks is closely related to the extrapolation error caused by the insufficient coverage of the relevant value range by the available training data. Its effect can be alleviated by providing well-defined input features reducing the total number of parameter combinations that have to be covered by the model. Using non-dimensional parameters according to [37], is recommended for gas path measurements since they collapse the engine performance to well-defined characteristics reducing the impact of environmental conditions [38]. These characteristics are mainly affected by controller settings such as bleed positions and airflow towards the cabin. Hence, whether or not a neural network can approximate the engine performance depends on the availability of sufficient data points with dedicated controller settings. An example of poor model accuracy related to insufficiently available controller settings during training is displayed in Figure 4. Since the data used for training the neural network were gathered during summer, data points with active anti-icing are scarce and the model's ability to correctly predict those operating regimes is limited. If the anti-icing is turned off, the approximations of the neural network match the in-flight measurements. However, over dedicated portions of the flight 315 s $\leq t \leq$ 2700 s and 4150 s $\leq t \leq$ 7100 s, the engine anti-icing is active, and the measurements are close to the upper prediction boundary of $\mu + 2\sigma$. If the tail anti-icing is turned on as well, deviations between the measurements and the neural network predictions increase further, surpassing the range of $\mu \pm 2\sigma$. The results lead to the conclusion that in order to prevent false positives originating from epistemic uncertainty, regions with high modeling uncertainty have to be identified and excluded.

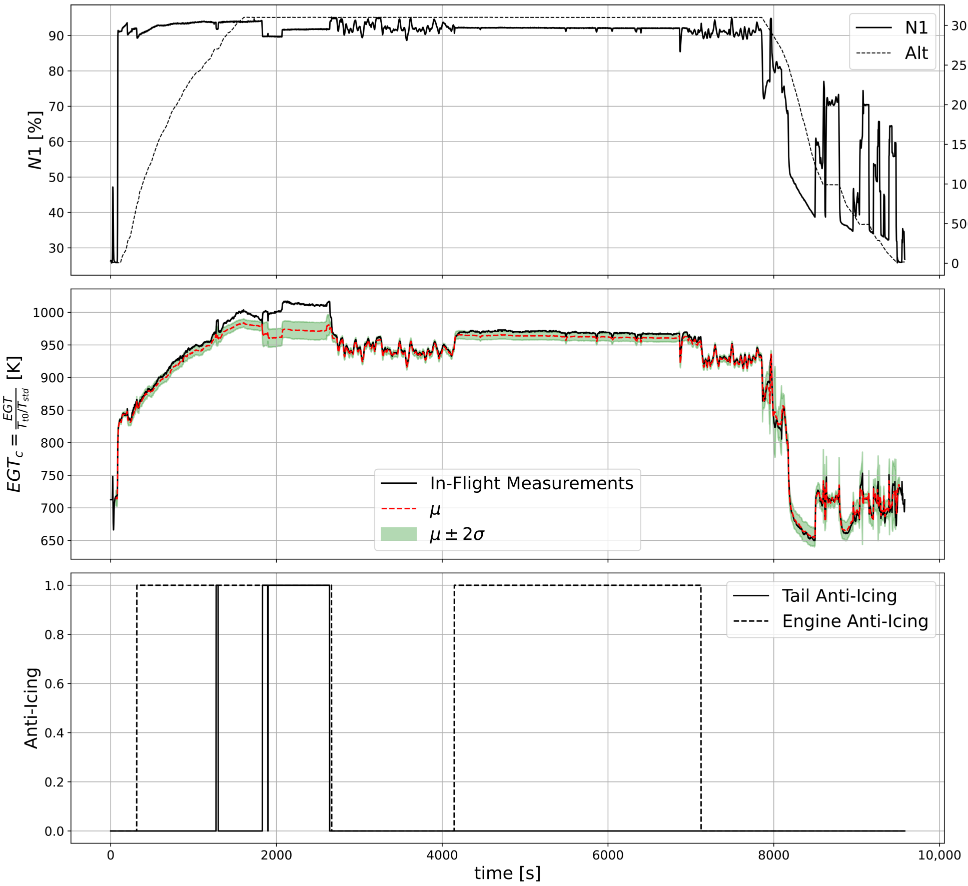

Figure 4. Approximation of the corrected exhaust gas temperature EGT_c for an example flight with varying anti-icing setting.

For the dataset examined in this article, there are a total of five controller settings available: engine anti-icing (*EAI*), tail anti-icing (*TAI*), wing anti-icing (*WAI*), bleed configuration and airflow towards the cabin (*Pack*). In order to quantify the availability of a sufficient number of data points within the training dataset, ensuring accurate model building, a confidence score $\mathcal{L}_{setting}$ is defined in Equation (6). The confidence score $\mathcal{L}_{setting}$ is based on the likelihood of occurrence of the controller settings $p_i(x(t))$, which are derived based on the dataset used for training the artificial neural network. In the proposed approach, the confidence score is computed separately for different flight phases PH to account for the impact of the operating conditions on the controller settings leading to conditional probabilities $p_i(x(t)|PH(t))$. Additionally, the probabilities are assumed to be statistically independent, neglecting the impact of different setting permutations.

$$\mathcal{L}_{setting}(t) = \prod p_i(x(t)|PH(t)) \\ i \in [EAI, TAI, WAI, Bleed, Pack]\,. \tag{6}$$

The resulting confidence score $\mathcal{L}_{setting}$ related to the previously shown flight is visualized in Figure 5. Since the confidence score is defined as the product of multiple probabilities leading to small values, the logarithmic confidence score $\mathcal{L}_{setting}$ is displayed here. The higher the logarithmic confidence score $\mathcal{L}_{setting}$, the more data points were available for training and the higher the model accuracy. Therefore, the confidence score $\mathcal{L}_{setting}$ can now be used to effectively exclude regions with high modeling uncertainty by defining an appropriate limit $\mathcal{L}_{lim}$. For the example flight, the timestamps with active tail-anti-icing around 1270 s $\leq t \leq$ 2640 s are characterized by a low confidence score $\mathcal{L}_{setting}$ resulting in high model uncertainty.

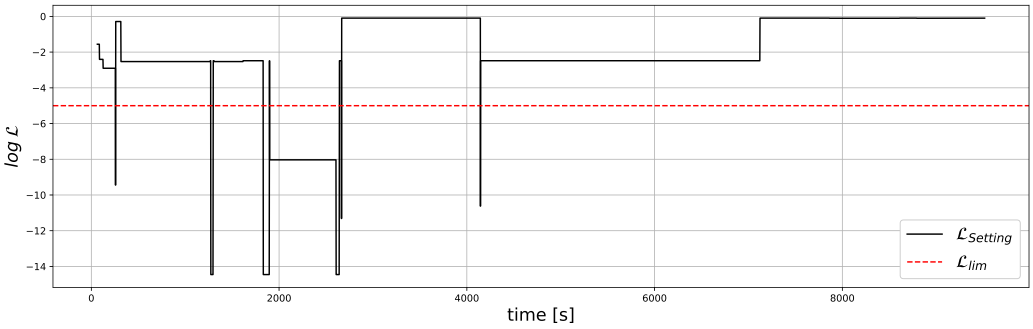

Figure 5. Corresponding confidence score $\mathcal{L}_{setting}$ for the flight displayed in Figure 4.

2.2. Fault Detection

Similar to [32], fault detection is based on the Mahalanobis Distance [39], defining the normalized distance of a test data point $\vec{y}_j$ from a probability density function

$$d_M(t) = (\vec{y}(t) - \vec{\mu}(t))^T \mathbf{\Sigma}(t)^{-1} (\vec{y}(t) - \vec{\mu}(t)) \tag{7}$$

The vector of means $\vec{\mu}$ and the correlation matrix $\mathbf{\Sigma}$ are the output of the neural network. The inverse of the correlation matrix $\mathbf{\Sigma}_j^{-1}$ in the definition of the Mahalanobis Distance d_M essentially weights the distances by the aleatoric uncertainty, ensuring that data points with high uncertainty are weighted less. This weighting directly reduces the risk of false positives in regions of high aleatoric uncertainty. Another advantage of the Mahalanobis Distance d_M is the definition of a single distance measure even for multivariate datasets. The availability of a single distance measure simplifies fault detection since only a single parameter has to be monitored.

The resulting Mahalanobis Distance d_M for a nominal example flight is visualized in Figure 6 alongside the flight profile. Especially for large transients, the artificial neural network has difficulties predicting the engine performance resulting in singular peaks in the Mahalanobis Distance d_M lasting only a few seconds. The fault detection scheme must account for these singular peaks to prevent false positives. In general, faults are considered to be persistent over a certain period of time, affecting the overall magnitude of the Mahalanobis Distance d_M. In order to avoid false positives triggered by singular events, the peaks in the Mahalanobis Distance d_M are removed by applying a Butterworth low-pass filter [40]. This low-pass filter ensures that only the magnitude of the Mahalanobis Distance d_M is considered for fault detection. The resulting Mahalanobis Distance d_M after applying the low-pass filter is additionally visualized in Figure 6.

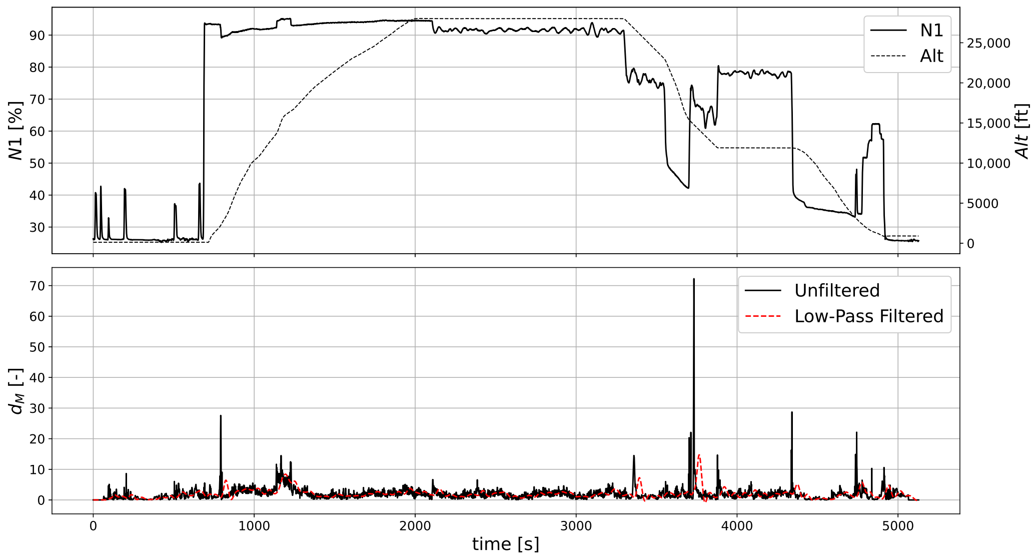

Figure 6. Resulting Mahalanobis Distance d_M for an example flight.

With the Mahalanobis Distance computed in every timestep, fully automatic fault detection can be conducted. The fault detection consists of the following three steps.

1. Identification of Regions with High Epistemic Uncertainty: Timestamps with high modeling uncertainty are identified and removed to ensure that only data points with high modeling accuracy are used for fault detection. Regions with high epistemic uncertainty are identified by defining a threshold on the confidence score $\mathcal{L}_{lim}$.
2. Outlier Detection: The detection of outliers indicating unusual system performance is based on the Mahalanobis Distance d_M. The corresponding data points are considered outliers if the Mahalanobis Distance d_M exceeds a predefined threshold $d_{M,lim}$.
3. Fault Detection: The total number of outliers is computed in the last step. Since there will always be a certain number of statistical outliers, a threshold n_{lim} on the total number of outliers is introduced. If the number of outliers detected exceeds this predefined threshold n_{lim}, the outliers are no longer considered statistical but systematic, indicating a fault.

2.3. Description of the Database

The proposed fault detection method is tested and trained with in-flight measurements of a commercially operated regional jet [41]. The dataset contains in-flight measurements of 35 aircraft covering a time period of three years. The data were anonymized, so there is no information about the aircraft or engine type. In general, the detection rates in engine condition monitoring depend highly on the model accuracy [42]. Since in flight-measurements vary due to production scatter and different degrees of degradation [43], only data of an individual engine serial number were extracted. Altogether 300 consecutive flights were extracted from the dataset. Nominal engine performance was ensured by comparing parallel mounted engines according to [44]. Since the in-flight measurements are acquired with different sampling rates, all measurements were first interpolated to a sampling rate of 1 Hz, the minimum sampling rate provided by most airlines [45]. Furthermore, only complete flights were extracted from the provided database.

The dataset covers more than 180 different parameters, mostly related to aircraft dynamics. Concerning gas-path measurements, only the measurements displayed in the cockpit $N1$, $N2$, EGT, Wf are provided. In order to limit the total number of input parameters to be processed by the artificial neural network, the dataset was manually filtered, extracting parameters that are considered to affect the performance of the aircraft engine. The resulting input and output parameters of the neural network are summarized in Table 1. In order to improve the training of the neural network [46], the discrete controller settings were normalized to $x \in [0, 1]$ and the continuous measurements were standardized to zero mean and a variance of one.

Table 1. Input and output parameters of the neural network.

Input Parameter	
Parameter	Description
$\theta = T_{t0}/T_{ISA}$	Correction factor temperature
$\delta = p_{t0}/p_{ISA}$	Correction factor pressure
$N1_c = N1/\sqrt{\theta}$	Corrected spool speed of the fan
MN	Mach-Number
BLV	Setting bleed extraction
$PACK$	Airflow towards the cabin
EAI	Setting engine-anti-icing
TAI	Setting tail-anti-icing
WAI	Setting wing-anti-icing
PH	Flight phase
Output Parameter	
Parameter	Description
$N2_c = N2/\sqrt{\theta}$	Corrected spool speed of the core
$EGT_c = EGT/\theta$	Corrected exhaust gas temperature
$Wf_c = Wf/(\sqrt{\theta}\delta)$	Corrected fuel flow

The provided dataset of full-flight data does not provide any information concerning potential faults. For a comprehensive investigation of the detection rates of the proposed fault detection scheme underlying various fault cases, synthetic datasets were generated by the superimposition of the in-flight measurements with measurement deviations generated utilizing a calibrated aircraft engine model of a regional jet. The fault cases were imposed by adjusting the capacities Q and efficiencies η of the engine components according to

$$\Delta Q = \left(Q - Q_{ref}\right) \Big/ Q_{ref} \tag{8}$$

$$\Delta\eta = \eta - \eta_{ref} \tag{9}$$

The scaling factors ΔQ and $\Delta\eta$ were chosen according to the OBIDICOTE test cases [47], which provide benchmark test cases for engine condition monitoring applications. The fault cases considered in this study and the corresponding scaling factors ΔQ and $\Delta\eta$ are summarized in Table 2.

Table 2. Definition of the OBIDICOTE test cases according to [47].

Label	**Fault Description**	
	ΔQ	$\Delta\eta$
a	$\Delta Q_{Fan} = -1.0\%$ $\Delta Q_{LPC} = -0.7\%$	$\Delta\eta_{Fan} = -0.5\%$ $\Delta\eta_{LPC} = -0.4\%$
b	–	$\Delta\eta_{Fan} = -1.0\%$
c	$\Delta Q_{HPC} = -1.0\%$	$\Delta\eta_{HPC} = -0.7\%$
d	–	$\Delta\eta_{HPC} = -1.0\%$

Table 2. *Cont.*

Label	Fault Description	
	ΔQ	$\Delta\eta$
e	$\Delta Q_{HPC} = -1.0\%$	–
f	$\Delta Q_{HPT} = +1.0\%$	–
g	$\Delta Q_{HPT} = -1.0\%$	$\Delta\eta_{HPT} = -1.0\%$
h	–	$\Delta\eta_{HPT} = -1.0\%$
i	–	$\Delta\eta_{LPT} = -1.0\%$
j	$\Delta Q_{LPT} = -1.0\%$	$\Delta\eta_{LPT} = -0.4\%$
k	$\Delta Q_{LPT} = -1.0\%$	–
l	$\Delta Q_{LPT} = +1.0\%$	$\Delta\eta_{LPT} = -0.6\%$

3. Results

3.1. Assessment of the Articifical Neural Network

Of the 300 flights extracted from the dataset of full-flight data, the first 100 consecutive flights were used for training and validating the artificial neural network. The flights within the training and validation dataset were randomly sampled, applying a ratio of 85%/15%, where the larger dataset was used for training the neural network. The remaining 200 flights are used to test the neural network and evaluate the detection rates. The training of the neural network was conducted for 1500 epochs utilizing Adam optimization [48] with a learning rate of $lr = 0.001$. Altogether 100 models were trained to account for randomness caused by the initialization of the neural network or the sampling of flights composing the training dataset. The neural network's architecture is constant for all models and was defined in advance by evaluating the loss functions for different architectures.

The output of the proposed neural network architecture for the corrected exhaust gas temperature EGT_c of an example flight in Figure 7 exemplifies the main advantage of utilizing uncertainty quantification. For neural networks without uncertainty quantification, e.g., trained on minimizing the mean squared error, the output will resemble the predictions for the mean exhaust gas temperature μ_{EGT}. While the approximated mean exhaust gas temperature μ_{EGT} can approximate the measured exhaust gas temperature EGT with high accuracy during climb and cruise, significant deviations are experienced during descent. These large deviations are mainly attributed to hysteresis in controller settings which are more dominant during descent and landing. Considering the engine's power setting, the shaft speed of the fan $N1$ is relatively stable during climb and cruise, while fast changes in $N1$ are dominant during descent and landing. Difficulties approximating the descent and landing are experienced for all flights within the training, validation, and test datasets, as can be seen considering the mean squared error mse in Table 3 and the mean standard deviation σ in Table 4. Since engine faults are identified by comparing the neural network's output with the in-flight measurements, such large deviations can lead to false positives if no uncertainty quantification is considered. On the other hand, the proposed neural networks with uncertainty quantification counteract the significant deviation by increasing the uncertainty, ultimately reducing the risk of false positives.

Table 3. Mean Squared Error mse for different flight phases.

	Climb			Cruise			Descent		
	mse_{EGT}	mse_{N2}	mse_{Wf}	mse_{EGT}	mse_{N2}	mse_{Wf}	mse_{EGT}	mse_{N2}	mse_{Wf}
Training	3.22 K	0.13%	0.64%	3.20 K	0.19%	0.74%	8.37 K	0.77%	2.36%
Validation	3.50 K	0.13%	0.69%	3.75 K	0.20%	0.82%	8.76 K	0.78%	2.44%
Testing	3.76 K	0.16%	0.78%	5.06 K	0.23%	0.98%	9.57 K	0.80%	2.55%

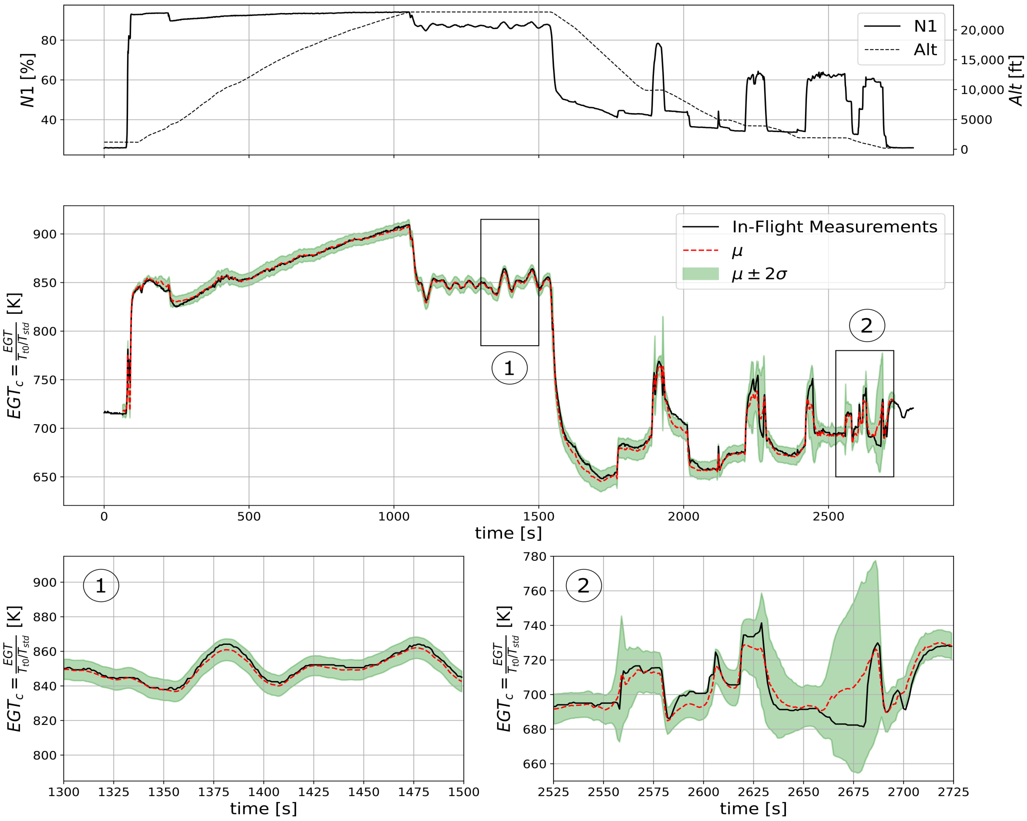

Figure 7. Approximation of the corrected exhaust gas temperature EGT_c for an example flight from the test dataset.

Table 4. Mean standard deviation σ for different flight phases.

	Climb			**Cruise**			**Descent**		
	σ_{EGT}	σ_{N2}	σ_{Wf}	σ_{EGT}	σ_{N2}	σ_{Wf}	σ_{EGT}	σ_{N2}	σ_{Wf}
Training	3.03 K	0.13%	0.66%	2.80 K	0.15%	0.73%	6.30 K	0.67%	2.63%
Validation	3.01 K	0.13%	0.66%	2.80 K	0.15%	0.73%	6.23 K	0.66%	2.60%
Testing	3.11 K	0.14%	0.68%	2.91 K	0.15%	0.74%	6.50 K	0.70%	2.71%

3.2. Detection Rates

The detection rates are evaluated by computing the true positive detection rates (TP) for the different fault cases and the false positive detection rates (FP) for nominal engine performance defined in Equations (10) and (11) [49].

$$TP_j = \frac{\text{Number of faults detected for fault case } j}{\text{Total number of flights with fault case } j} \tag{10}$$

$$FP = \frac{\text{Number of faults detected for nominal flights}}{\text{Total number of nominal flights}} \,. \tag{11}$$

The proposed fault detection algorithm features three thresholds directly affecting its sensitivity for fault detection: the limit on the confidence score $\mathcal{L}_{lim}$ ensuring model quality, the limit on the Mahalanobis Distance $d_{M,lim}$ used for detecting outliers, and the total number of outliers tolerated until fault detection n_{lim}. To determine the optimal combination of thresholds, a grid search was performed, discretizing the limits and searching for parameter combinations that achieve maximum average true positive detection rates $\overline{TP}$ for predefined thresholds on the maximum allowable false positive detection rates

$FP \leq FP_{lim}$. Here, the average true positive detection rate $\overline{TP}$ was computed, taking into account the true positive detection rates TP_j of the individual OBIDICOTE test cases.

The resulting average true positive detection rates $\overline{TP}$ and the corresponding limits on the outliers tolerated until fault detection n_{lim} for the algorithms with and without additional estimation of the epistemic uncertainty are visualized in Figure 8. Since all 100 trained models were evaluated, the results are statistically evaluated and visualized as box-plots. The results clearly show the advantage of performing an additional estimation of the epistemic uncertainty. The results with the additional estimation of the epistemic uncertainty require smaller limits on the outliers tolerated until fault detection n_{lim} while achieving higher average true positive detection rates $\overline{TP}$. The difference between the two methods becomes more pronounced when requiring small false positive detection rates FP. Considering $FP \leq 0.5\%$, the presented method improves the average true positive detection rate $\overline{TP}$ by a factor of 2.8 compared to the method only accounting for the aleatoric uncertainty. Furthermore, the number of outliers tolerated n_{lim} and consequently the response time can be improved by approximately 6.9. Especially when requiring low false positive detection rates FP, the resulting true positive detection rates TP are too low for operational application when only the aleatoric uncertainty is approximated. For the presented use case with a sampling rate of 1 Hz, a period of time with on average 7.9 min of faulty engine performance during a 75 min flight is required until fault detection.

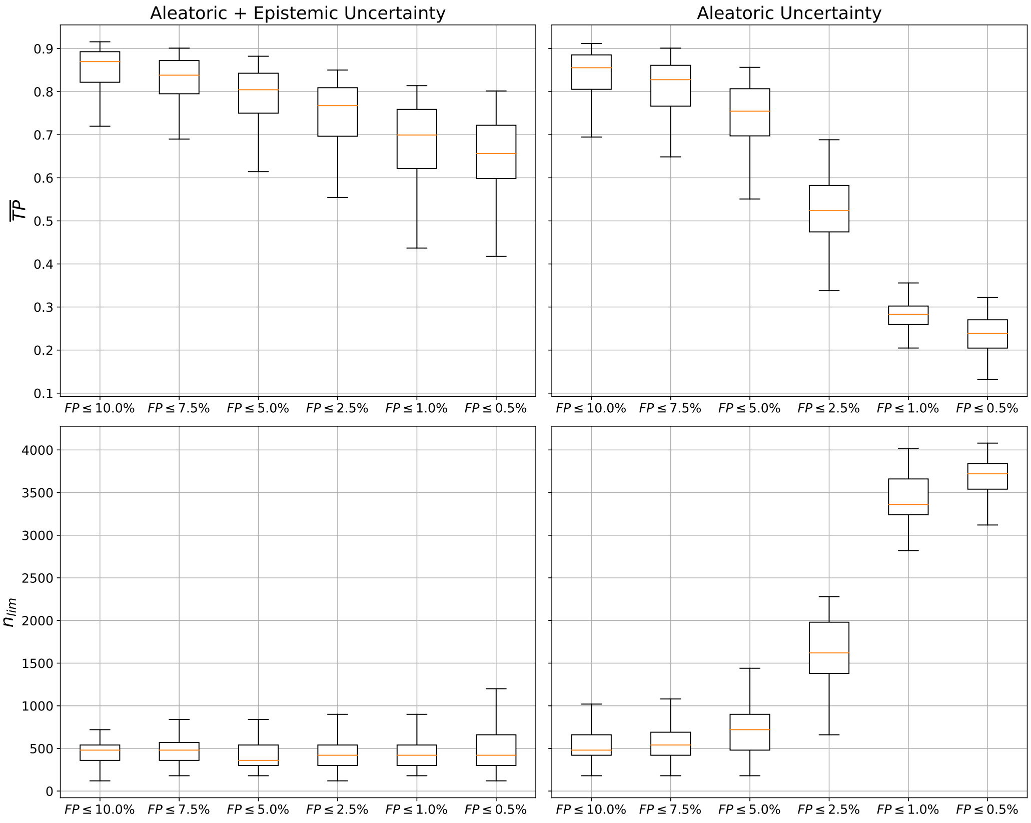

Figure 8. Comparison of the average true positive detection rate $\overline{TP}$ and number of outliers tolerated until fault detection n_{lim}.

Considering the median true positive detection rate TP for the different fault cases with aleatoric and epistemic uncertainty quantification summarized in Table 5 reveals that the poor average detection rates $\overline{TP}$ mainly result from difficulties identifying fault case f. Since only minimum instrumentation is provided for the available data set, observability issues exist for certain fault cases. Incorporating more sensors within the fault detection algorithm can improve the detection rates. Due to its modular architecture, the proposed fault detection approach can be easily extended for different measurement suits.

Table 5. Resulting detection rates for the OBIDICOTE test cases a–l (Table 2) for aleatoric and epistemic uncertainty quantification.

	TP											
	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**
$FP \leq 10.0\%$	97.7%	89.0%	84.5%	98.3%	87.7%	24.7%	98.3%	99.7%	97.0%	87.0%	91.7%	97.3%
$FP \leq 7.5\%$	95.7%	86.7%	81.3%	96.7%	84.0%	19.7%	96.7%	97.7%	95.3%	83.7%	89.0%	95.7%
$FP \leq 5.0\%$	94.0%	82.0%	74.3%	95.0%	79.3%	14.7%	95.0%	96.0%	94.3%	77.3%	85.7%	94.3%
$FP \leq 2.5\%$	92.3%	75.7%	67.3%	93.7%	71.3%	9.3%	93.7%	95.3%	93.3%	71.7%	81.0%	93.7%
$FP \leq 1.0\%$	88.3%	62.3%	52.3%	90.3%	57.0%	4.0%	89.7%	94.0%	88.0%	62.3%	74.0%	90.3%
$FP \leq 0.5\%$	86.3%	55.3%	47.0%	87.7%	49.7%	3.3%	87.3%	92.3%	85.7%	57.0%	69.3%	88.0%

3.3. Sensitivity Study: Fault Initiation

The results presented in the previous section were derived by initiating the fault right at the start of the time series $t = 0$. Additional examinations were conducted to quantify the sensitivity of the detection rates concerning the point in time when the faults are initiated. For this sensitivity study, the thresholds on the confidence score $\mathcal{L}_{lim}$, Mahalanobis Distance $d_{M,lim}$, and the total number of outliers tolerated until fault detection n_{lim} retained in the previous section are kept constant. The faults are initialized relative to the total flight length.

The median average true positive detection rates $\overline{TP}$ for different relative fault initiation times t_{init} are displayed in Figure 9. The results show compromising detection rates if the fault happens later during the flight. For example, if a fault is initiated within the last 25% of the flight, the maximum achievable average detection rates are less than 35%. The decreased performance of the fault detection approach in later flight phases is mainly attributed to the increased uncertainty experienced during descent and landing, already described in Section 3.1. Correspondingly, only faults strongly affecting the measurements can be detected. In the worst case, the fault can not be detected within the current flight. However, the chances are high that the fault can be detected within the next flight, which is still an improvement compared to current state-of-the-art methods [5,7,8].

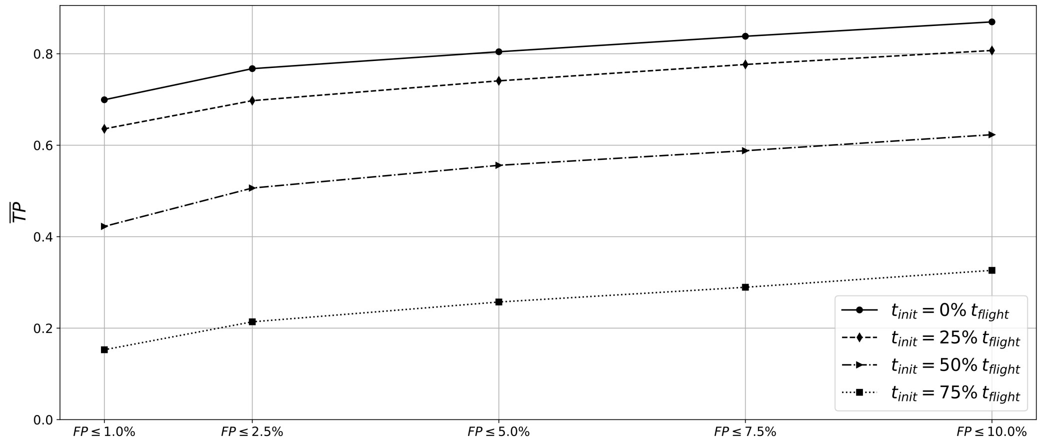

Figure 9. Sensitivity of the median true positive detection rate $\overline{TP}$ with respect to the fault initiation t_{init}.

4. Discussion

This paper presents a novel approach for estimating the aleatoric and epistemic uncertainty in data-driven engine fault detection. The algorithm can detect arbitrary faults requiring only datasets representing nominal engine performance. All tests conducted were based on in-flight data of a commercially operated regional jet, ensuring real changes in environmental conditions and controller settings. Compared to alternative approaches only accounting for the aleatoric uncertainty, the presented approach results in improved detection rates and faster response times. Especially if low false positive detection rates are required, methods based on only the aleatoric uncertainty lead to too low true positive detection rates unsuitable for operational application. Various fault cases could be detected within a single flight removing the latency of current state-of-the-art fault detection based on steady-state snapshots. For the tests, only minimal instrumentation was provided. Fault detection can potentially be further enhanced by providing additional sensors to improve the observability of the engine.

In the presented use case, the engine model was trained based on datasets of an individual engine to avoid the impact of production scatter and account for engine degradation. To ensure fast coverage of an engine within condition monitoring, the dataset used for training the model covers only a short period of time, limiting the diversity of training data and increasing the epistemic uncertainty.

Author Contributions: Conceptualization, M.W., S.S., D.B., C.K. and J.M.; methodology, M.W. and S.S.; software, M.W.; validation, M.W., S.S., D.B., C.K. and J.M.; writing—original draft preparation, M.W. and S.S.; writing—review and editing, S.S., D.B., C.K. and J.M.; visualization, M.W.; supervision, S.S.; and funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the German Federal Ministry of Economic Affairs and Energy (BMWI) within the LUFO-V project OPsTIMAL (grant number 20X1711D).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://c3.ndc.nasa.gov/dashlink/projects/85/ (accessed on 25 May 2020).

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

The following nomenclature are used in this manuscript:

Symbols

Alt	Altitude
BLV	Bleed Setting
d_M	Mahalanobis Distance
$d_{M,lim}$	Threshold for the Mahalanobis Distance
EGT	Exhaust Gas Temperature
EAI	Engine Anti-Icing
FP	False Positive Detection Rate
l	Flight Length
lr	Learning Rate
MN	Mach-Number
mse	Mean Squared Error
$\mathcal{L}_{lim}$	Threshold for the Confidence Score
$\mathcal{L}_{setting}$	Confidence Score
n_{lim}	Threshold for the Number of Outliers Tolerated
$N1$	Shaft Speed Fan
$N2$	Shaft Speed Core
$\mathcal{NLL}$	Negative Log-Likelihood
p	Pressure
$PACK$	Airflow towards the Cabin
PH	Flight Phase
Q	Capacity

T	Temperature
t	Time
t_{init}	Time Until Fault Initiation
TAI	Tail Anti-Icing
TP	True Positive Detection Rate
WAI	Wing Anti-Icing
Wf	Fuel Flow
x	Input-Parameter
y	Measurement
Δ	Deviation
δ	Correction Factor Pressure
Σ	Correlation Matrix
σ	Standard Deviation
μ	Mean Value
η	Efficiency
Θ	Parameter Set
θ	Correction Factor Temperature

Superscripts and Subscripts

$\bar{f}$	Averaged Value
c	ISA-Corrected Value
ISA	ISA Reference Value

Acronyms

CC	Combustion Chamber
HPC	High Pressure Compressor
HPT	High Pressure Turbine
LPC	Low Pressure Compressor
LPT	Low Pressure Turbine

References

1. IATA. *Airline Maintenance Cost Executive Commentary*; Technical Report; IATA: Montreal, QC, Canada, 2016.
2. Fentaye, A.; Baheta, A.T.; Gilani, S.I.; Kyprianidis, K.G. A Review on Gas Turbine Gas-Path Diagnostics: State-of-the-Art Methods, Challenges and Opportunities. *Aerospace* **2019**, *6*, 83. [CrossRef]
3. Fentaye, A.D.; Zaccaria, V.; Kyprianidis, K. Aircraft Engine Performance Monitoring and Diagnostics Based on Deep Convolutional Neural Networks. *Machines* **2021**, *9*, 337. [CrossRef]
4. Fentaye, A.D.; Ul-Haq Gilani, S.I.; Baheta, A.T.; Li, Y.G. Performance-based fault diagnosis of a gas turbine engine using an integrated support vector machine and artificial neural network method. *Proc. Inst. Mech. Eng. Part J. Power Energy* **2019**, *233*, 786–802. [CrossRef]
5. Pérez-Ruiz, J.L.; Tang, Y.; Loboda, I. Aircraft Engine Gas-Path Monitoring and Diagnostics Framework Based on a Hybrid Fault Recognition Approach. *Aerospace* **2021**, *8*, 232. [CrossRef]
6. Lipowsky, H.; Staudacher, S.; Bauer, M.; Schmidt, K.J. Application of Bayesian Forecasting to Change Detection and Prognosis of Gas Turbine Performance. *J. Eng. Gas Turbines Power* **2010**, *132*, 1–8. [CrossRef]
7. Koskoletos, O.A.; Aretakis, N.; Alexiou, A.; Romesis, C.; Mathioudakis, K. Evaluation of Aircraft Engine Diagnostic Methods Through ProDiMES. In Proceedings of the ASME Turbo Expo 2018: Turbomachinery Technical Conference and Exposition (GT2018), Oslo, Norway, 11–15 June 2018; ASME: New York, NY, USA, 2018; p. V006T05A023. [CrossRef]
8. Loboda, I.; Pérez-Ruiz, J.L.; Yepifanov, S. A Benchmarking Analysis of a Data-Driven Gas Turbine Diagnostic Approach. In Proceedings of the ASME Turbo Expo 2018: Turbomachinery Technical Conference and Exposition (GT2018), Oslo, Norway, 11–15 June 2018; ASME: New York, NY, USA, 2018; p. V006T05A027. [CrossRef]
9. Badea, V.E.; Zamfiroiu, A.; Boncea, R. Big Data in the Aerospace Industry. *Inform. Econ.* **2018**, *22*, 17–24. [CrossRef]
10. Volponi, A.J.; Tang, L. Improved Engine Health Monitoring Using Full Flight Data and Companion Engine Information. *SAE Int. J. Aerosp.* **2016**, *9*, 91–102. [CrossRef]
11. Tang, L.; Volponi, A.J.; Prihar, E. Extending engine gas path analysis using full flight data. *Proc. ASME Turbo Expo* **2019**, *6*, 1–11. [CrossRef]
12. Weiss, M.; Staudacher, S.; Becchio, D.; Keller, C.; Mathes, J. Steady-State Fault Detection with Full-Flight Data. *Machines* **2022**, *10*, 140. [CrossRef]
13. Dai, X.; Gao, Z. From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Trans. Ind. Inform.* **2013**, *9*, 2226–2238. [CrossRef]

14. Zhao, F.; Dasgupta, A.; Yuan, C.; Chakraborty, A. Multi-Level Neural Network Based Gas Turbine Modeling. In Proceedings of the ASME Turbo Expo 2018: Turbomachinery Technical Conference and Exposition (GT2018), Oslo, Norway, 11–15 June 2018.
15. Bai, M.; Liu, J.; Ma, Y.; Zhao, X.; Long, Z.; Yu, D. Long Short-Term Memory Network-Based Normal Pattern Group for Fault Detection of Three-Shaft Marine Gas Turbine. *Energies* **2020**, *14*, 13. [CrossRef]
16. Pogorelov, G.; Kulikov, G.; Abdulnagimov, A.; Badamshin, B. Application of Neural Network Technology and High-performance Computing for Identification and Real-time Hardware-in-the-loop Simulation of Gas Turbine Engines. *Procedia Eng.* **2017**, *176*, 402–408. [CrossRef]
17. Goyal, V.; Xu, M.; Kapat, J.; Vesely, L. Prediction of gas turbine performance using machine learning methods. *Proc. ASME Turbo Expo* **2020**, *6*, 1–11. [CrossRef]
18. Castillo, I.G.; Loboda, I.; Pérez Ruiz, J.L. Data-Driven Models for Gas Turbine Online Diagnosis. *Machines* **2021**, *9*, 372. [CrossRef]
19. Loboda, I.; Feldshteyn, Y. Polynomials and neural networks for gas turbine monitoring: A comparative study. *Int. J. Turbo Jet Engines* **2011**, *28*, 227–236. [CrossRef]
20. Goyal, V.; Xu, M.; Kapat, J. Statistical modeling in failure detection in gas turbines. In Proceedings of the AIAA Propulsion and Energy Forum and Exposition, Indianapolis, IN, USA, 19–22 August 2019; pp. 1–10. [CrossRef]
21. Zhang, C.; Wang, N. Aero-Engine Condition Monitoring Based on Support Vector Machine. *Phys. Procedia* **2012**, *24*, 1546–1552. [CrossRef]
22. Michelmore, R.; Wicker, M.; Laurenti, L.; Cardelli, L.; Gal, Y.; Kwiatkowska, M. Uncertainty Quantification with Statistical Guarantees in End-to-End Autonomous Driving Control. In Proceedings of the IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 7344–7350. [CrossRef]
23. Kompa, B.; Snoek, J.; Beam, A.L. Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digit. Med.* **2021**, *4*, 1–6. [CrossRef]
24. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **2021**, *110*, 457–506. [CrossRef]
25. Haley, P.; Soloway, D. Extrapolation limitations of multilayer feedforward neural networks. In Proceedings of the IJCNN International Joint Conference on Neural Networks, Baltimore, MD, USA, 7–11 June 1992; Volume 4, pp. 25–30. [CrossRef]
26. McCartney, M.; Haeringer, M.; Polifke, W. Comparison of Machine Learning Algorithms in the Interpolation and Extrapolation of Flame Describing Functions. *J. Eng. Gas Turbines Power* **2020**, *142*, 061009. [CrossRef]
27. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Adv. Neural Inf. Process. Syst.* **2016**, *2017*, 6403–6414.
28. Bishop, C. Novelty detection and neural network validation. *IEE Proc. Vision Image Signal Process.* **1994**, *141*, 217. [CrossRef]
29. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 20–22 June 2016; Volume 3, pp. 1651–1660.
30. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Networks. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015; Volume 2, pp. 1613–1622. [CrossRef]
31. Likas, A. Probability density estimation using artificial neural networks. *Comput. Phys. Commun.* **2001**, *135*, 167–175. [CrossRef]
32. Hartwell, A.; Montana, F.; Jacobs, W.; Kadirkamanathan, V.; Mills, A.R.; Clark, T. In-flight Novelty Detection with Convolutional Neural Networks. *arXiv* **2021**, arXiv:2112.03765.
33. Putz, A.; Staudacher, S.; Koch, C.; Brandes, T. Jet Engine Gas Path Analysis Based on Takeoff Performance Snapshots. *J. Eng. Gas Turbines Power* **2017**, *139*, 111201. [CrossRef]
34. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
35. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
36. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
37. Volponi, A.J. Gas Turbine Parameter Corrections. *J. Eng. Gas Turbines Power* **1999**, *121*, 613–621. [CrossRef]
38. Walsh, P.P.; Fletcher, P. *Gas Turbine Performance*, 2nd ed.; Wiley Blackwell: Oxford, UK, 2004. [CrossRef]
39. Mahalanobis, P. On the generalized distance in statistics. In *National Institute of Science of India*; National Institute of Science of India: Calcutta, India, 1936; pp. 49–55.
40. Butterworth, S. On the theory of filter amplifiers. *Wirel. Eng.* **1930**, *7*, 536–541.
41. Matthews, B.; Oza, N. NASA—Sample Flight Data. 2012. Available online: https://c3.ndc.nasa.gov/dashlink/projects/85/ (accessed on 25 May 2020).
42. Köhli, R. Untersuchungen zum Einfluss der Modellbildung auf das Trend Monitoring von Fluggasturbinen. Ph.D Thesis, Universität Stuttgart, Stuttgart, Germany, 2016.
43. Spieler, S.; Staudacher, S.; Fiola, R.; Sahm, P.; Weißschuh, M. Probabilistic engine performance scatter and deterioration modeling. *J. Eng. Gas Turbines Power* **2008**, *130*, 042507. [CrossRef]
44. Babbar, A.; Ortiz, E.M.; Syrmos, V.L.; Arita, M.M. Advanced diagnostics and prognostics for engine health monitoring. In Proceedings of the IEEE Aerospace Conference Proceedings, Big Sky, MN, USA, 7–14 March 2009; pp. 1–10. [CrossRef]

45. Sheridan, K.; Puranik, T.G.; Mangortey, E.; Pinon-Fischer, O.J.; Kirby, M.; Mavris, D.N. An Application of DBSCAN Clustering for Flight Anomaly Detection During the Approach Phase. In Proceedings of the AIAA Scitech 2020 Forum, Orlando, FL, USA, 6–10 January 2020; American Institute of Aeronautics and Astronautics: Reston, VI, USA, 2020. [CrossRef]
46. da Silva, I.N.; Hernane Spatti, D.; Andrade Flauzino, R.; Liboni, L.H.B.; dos Reis Alves, S.F. *Artificial Neural Networks*; Springer International Publishing: Cham, Switzerland, 2017; p. 307. [CrossRef]
47. Curnock, B. *Obidicote Project—Work Package 4: Steady-State Test Cases*; Technical Report DNS62433; Rolls-Royce Plc: Manchester, UK, 2000.
48. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
49. Simon, D.L.; Bird, J.; Davison, C.; Volponi, A.; Iverson, R.E. Benchmarking Gas Path Diagnostic Methods: A Public Approach. In Proceedings of the ASME Turbo Expo 2008: Power for Land, Sea, and Air, Berlin, Germany, 9–13 June 2008; Volume 2, pp. 325–336. [CrossRef]