aerospace

Article

Exhaust Gas Temperature Prediction of Aero-Engine via Enhanced Scale-Aware Efficient Transformer

Sijie Liu, Nan Zhou, Chenchen Song, Geng Chen and Yafeng Wu *

School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China;
sijieliu_123@mail.nwpu.edu.cn (S.L.); xiguanan@mail.nwpu.edu.cn (N.Z.); 2022201912@mail.nwpu.edu.cn (C.S.);
geng.chen.cs@gmail.com (G.C.)
* Correspondence: yfwu@nwpu.edu.cn; Tel.: +86-029-88431112

Abstract: This research introduces the Enhanced Scale-Aware efficient Transformer (ESAE-Transformer), a novel and advanced model dedicated to predicting Exhaust Gas Temperature (EGT). The ESAE-Transformer merges the Multi-Head ProbSparse Attention mechanism with the established Transformer architecture, significantly optimizing computational efficiency and effectively discerning key temporal patterns. The incorporation of the Multi-Scale Feature Aggregation Module (MSFAM) further refines 2 s input and output timeframe. A detailed investigation into the feature dimensionality was undertaken, leading to an optimized configuration of the model, thereby improving its overall performance. The efficacy of the ESAE-Transformer was rigorously evaluated through an exhaustive ablation study, focusing on the contribution of each constituent module. The findings showcase a mean absolute prediction error of 3.47° R, demonstrating strong alignment with real-world environmental scenarios and confirming the model's accuracy and relevance. The ESAE-Transformer not only excels in predictive accuracy but also sheds light on the underlying physical processes, thus enhancing its practical application in real-world settings. The model stands out as a robust tool for critical parameter prediction in aero-engine systems, paving the way for future advancements in engine prognostics and diagnostics.

Keywords: exhaust gas temperature prediction; ESAE-Transformer; Multi-Scale Feature Aggregation

check for updates

Citation: Liu, S.; Zhou, N.; Song, C.; Chen, G.; Wu, Y. Exhaust Gas Temperature Prediction of Aero-Engine via Enhanced Scale-Aware Efficient Transformer. *Aerospace* **2024**, *11*, 138. https://doi.org/10.3390/aerospace11020138

Academic Editor: Olivia J. Pinon Fischer

Received: 18 December 2023
Revised: 23 January 2024
Accepted: 29 January 2024
Published: 5 February 2024

1. Introduction

Aero-engines are of most importance in ensuring the proper functioning of aircrafts, given their intricate design and the possibility of catastrophic malfunctions. These engines endure extended periods of operation under severe environmental conditions, including elevated temperatures, pressures, velocities, vibrations, and loads [1,2]. Therefore, it is crucial to guarantee the dependability of aero-engines in accordance with rigorous safety protocols by continuously monitoring and forecasting the fundamental parameters of the aero-engine [3–5].

Exhaust gas temperature (EGT) refers to the temperature of the gas that exits the turbine unit, which is regarded as one of the most crucial structural and operational metrics that demonstrate the performance and efficiency of gas turbine engines [1,6]. Elevated EGT can result in significant faults and diminish the engine's longevity. It is imperative to monitor the EGT during take-off and strive to keep it at a minimum. This is crucial since the EGT reaches its highest point during take-off, and exceeding the usual limits of EGT can lead to engine component failure. From a structural standpoint, accurately predicting the EGT has several benefits, including improved reliability, availability, and engine life extension, as well as reduced operation and maintenance costs [7]. In aeronautic applications, turbine engines are used to provide the necessary thrust or power throughout different flight phases by either increasing or decreasing the velocity of the air passing through the engines. When dealing with various take-off situations, it is important to find a balance between generating sufficient thrust while keeping the EGT relatively low. Evaluating the EGT level is critical for assessing both the structural and operational aspects, as mentioned before.

For the predicting the EGT of an aero-engine based on real flight data, the current prediction methods are widely classified into two categories: the model-based method and the data-driven method. The model-based approach relies on the precise physical models of the system, which are combined with mathematical and physical models that describe the dynamic performance of the aeroengine. The model-based method is constrained in its applicability due to the need for accurate modeling of the dynamics of mechanical systems or components [8]. Nevertheless, it is impossible to achieve accurate modeling of intricate systems, even for individuals with specialized knowledge in the domain. Data-driven approaches, as opposed to model-based approaches, offer the benefit of being easier to implement due to their lack of reliance on prior professional competence. Hence, the utilization of data-driven approaches is more common in modern industrial practices [9].

Based on data-driven approaches, data-driven approaches can be categorized into three types: statistical, machine learning, and deep learning [10]. The statistical methods commonly used for industrial prediction problems include the autoregressive (AR) model, the autoregressive integrated moving average (ARIMA) model, random forest (RF), and Kalman filters (KF). Given the quickly changing nature of the EGT temporal data, it is evident that traditional statistical methods, which are designed for linear stationary data without differences, are not suitable for accurately predicting EGT [11]. With the boom in the development of machine learning and deep learning techniques, as well as the progress of sensor technology and real-time databases, the data-driven prediction of engine state parameters has attracted wide attention from academia and industry [12]. In their study on EGT prediction, Wang et al. [13] established basic frameworks and employed several common machine learning methods. The analysis included the use of the Generalized Regression Neural Network (GRNN) network [14], the Radial Basis Function (RBF) network [15], Support Vector Regression (SVR) [16], and Random Forest (RF) [17]. An EGT prediction approach utilizing a long short-term memory (LSTM) network was proposed by Ullah et al. [18]. The features provided as input were identified as a real-time series. Other efforts are built around the NARX model, which is a kind of recurrent neural network. This model is able to capture the intricate dynamics of complex systems, such as gas turbines, and can be integrated with different types of neural networks. Asgari et al. [19] have developed NARX models for a robust single-shaft gas turbine. The findings demonstrated the utility of NARX models in predicting the dynamic response of gas turbines. Pham et al. [20] proposed an enhancement to the hybrid NARX model and autoregressive moving average (ARMA) model for the long-term prediction of machine state using vibration data. Ma et al. [21] use an FAE-LSTM, a feature attention mechanism-enhanced (FAE) LSTM, to build a NARX model. This model uses EGT-correlated condition features and gas path measurement factors to identify the aircraft engine. The Moving Average (MA) model uses a basic LSTM model to increase the difference between the observed EGT and the NARX model's anticipated EGT. The endeavour of the aforementioned is to create a real-time dynamic prediction model that can accurately capture the changing reaction of the EGT in different operational states and challenging work settings.

Recently, the self-attention-based Transformer model has been widely used in service industrial areas, including NLP, computer vision, and time series prediction. Self-attention, also known as intra-attention, may be seen as an attentional process [22]. Natural language processing mostly utilizes the self-attention-based transformer, as proposed by [23]. The transformer-based model is primarily used for the important task of predicting the remaining useful life (RUL) of aeroengines. Zhang et al. integrated the BiGRU encoder and the Transformer decoder in order to develop a network structure for predicting the RUL of turbofan engines [10]. In their study, Liu et al. [24] developed a double attention network that incorporated a multi-head attention module and a 2-D CNN-based channel attention enhancement module. The primary objective of this network was to enhance the accuracy of remaining useful life (RUL) predictions in four different working scenarios.

Given the significant progress made in relevant research, there are still some outstanding obstacles that need more exploration. Unlike the forecast of RUL, which mainly considers the evaluation of the whole lifespan, the prediction of EGT requires a more detailed technique that specifically highlights the time-based features present in actual flight data. Currently, prevalent methods for predicting EGT encompass both physical models and machine learning predictions. Physical models typically involve cylinder combustion models, heat transfer models, and exhaust models. However, the intricate phenomena occurring in the later stages of combustion such as boundary layer effects, uneven fuel distribution, heat conduction, and other factors can pose challenges for traditional physical models in accurately forecasting EGT. Alternatively, by adopting data-driven prediction models and employing machine learning algorithms to simulate combustion, heat transfer, and exhaust cooling processes, the complexity associated with EGT prediction can be mitigated. This approach has the potential to assist or even replace traditional physical prediction models. In order to efficiently track the decline in system performance and accurately capture the important time-related characteristics, we propose the adoption of a transformer-based model. To address these issues, we provide a novel approach Enhanced Scale-Aware efficient Transformer (ESAE-Transformer): a Transformer model with Multi-Head ProbSparse Self-Attention (MHPSA) and a Multi-Scale Feature Aggregation module (MSFAM). This model is a comprehensive framework that consists of an encoder and a decoder. The encoder and decoder are upgraded with MHPSA to effectively capture important operational variations and environmental changes and reduce the compute complexity. Simultaneously, the Multi-Scale Feature Aggregation module (MSFAM) is used to enhance the high-dimensional encoded feature space, hence expanding the range of information that can be captured. This design is expected to greatly improve the precision and effectiveness of EGT forecasts. The main contribution of this work can be summarized as follows:

- We recommend the utilization of a specialized model designed for predicting EGT, and this model is built upon the innovative transformer architecture. To our knowledge, this groundbreaking initiative represents the first successful effort to tailor the transformer design specifically for RUL in the context of aero-engines.
- The encoder and decoder models leverage an MHPSA mechanism, strategically designed to efficiently decrease temporal complexity and optimize memory utilization. This innovative approach introduces the concept of selective attention, empowering the model to concentrate on the most informative segments. This not only diminishes noise but also prioritizes critical temporal dynamics, enhancing the overall effectiveness of the system.
- The implementation of an MSFAM is purposefully crafted to delve into temporal features within a profoundly nonlinear dimensional space. Its primary function is to broaden the receptive field of the prediction model, thereby enhancing the model's proficiency in effectively processing and amalgamating implicit information across extended sequences or time periods. This strategic design significantly improves the model's capacity to capture and leverage nuanced temporal dynamics for more robust predictions.
- To assess the suggested approach, we conducted evaluations on two fronts. Firstly, we compared the root mean square difference and absolute mean error of predicted results against actual results, varying the dimension of the hidden layer while adjusting the length of the time series input to the model. Optimal performance was observed when the input length was 2 s, and the model dimension was set to 128. Secondly, across different input lengths, we compared our proposed model with contemporary time series prediction models like ANN, LSTM, GRU, and Transformer. The experimental findings revealed that our proposed model outperforms current popular time series prediction models based on the same evaluation criteria.

The remainder of the paper is organised as follows: The core concept of the vanilla Transformer is presented in Section 2. The proposed method is described in depth in Section 3. The dataset description and experimental parameters are provided in Section 4. Section 5 presents the outcomes of the experiment and provides a thorough analysis. The entire essay is summarized in Section 6.

2. The Fundamental Principle of the Transformer

The Transformer model, initially introduced in the influential paper titled "Attention Is All You Need", produced by Vaswani et al. in 2017, has significantly influenced the field of natural language processing (NLP) [23]. The overall structure of the Transformer is shown in Figure 1. The model utilizes a methodology known as self-attention or scaled dot-product attention, hence circumventing the inclusion of recurrent layers commonly found in prior sequence-to-sequence models. This particular design decision offers advantages in terms of parallelization and reduced training times.

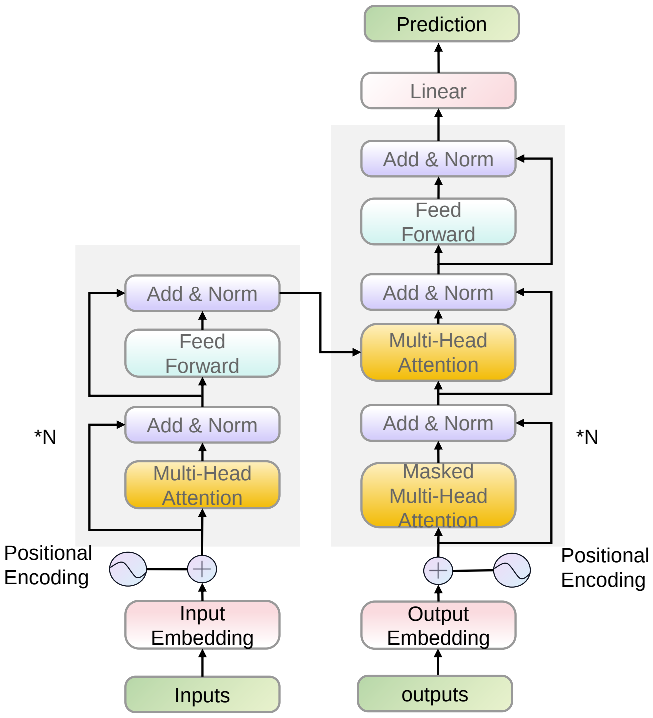


Figure 1. The overall structure of the Transformer [23] and * represents the stacked encoder and decoder modules.

The Transformer model is characterized by an encoder–decoder structure. The decoder shares important components with the encoder, including Position data embedding, Multi-Head Self-Attention, and Point-Wise Feed-Forward Network. Each sub-layer in the encoder and decoder has a residual connection around it, followed by layer normalization. This design facilitates the flow of gradients during training, making it easier to train deep networks. In both the encoder and decoder, each sub-layer is equipped with a residual connection, which is then followed by layer normalization. This particular architecture enhances the smooth transition of gradients during the training process, hence simplifying the training of deep neural networks.

2.1. Position Data Embedding

Positional encoding is a technique that imparts the model with knowledge regarding the relative or absolute position of the tokens inside the sequence. The positional encodings and embeddings possess equivalent dimensions, denoted as d_{model}, enabling their summation. This implies that the embedding of each token is modified by the addition of a vector that signifies the token's positional information inside the sequence. The positional encodings use sine and cosine functions of different frequencies.

For each position *pos* and each dimension *i* of the d_{model} token embedding, the positional embedding $PE_{(pos,i)}$ is defined as (1):

$$\begin{cases} \text{PE}_{(pos,2i)} = \sin\left(\dfrac{pos}{10000^{2i/d_{\text{model}}}}\right) \\ \text{PE}_{(pos,2i+1)} = \cos\left(\dfrac{pos}{10000^{2i/d_{\text{model}}}}\right) \end{cases} \tag{1}$$

where $\text{PE}_{(pos,2i)}$ and $\text{PE}_{(pos,2i+1)}$ denote the positional embedding for a given position and dimension i in the embedding of the model. The sine function is used for even indices $2i$, while the odd indices $2i+1$ use the cosine function. The $10000^{2i/d_{\text{model}}}$ term provides scaling that allows the model to learn to attend by relative positions more easily.

By adding positional encoding to the input embeddings, the Transformer becomes capable of considering the order of the sequence, which is critical for understanding language and other sequence-based data.

2.2. Multi-Head Self-Attention

The incorporation of Multi-Head Self-Attention (MHSA) in Transformer models enables the model to collectively attend to input originating from distinct representation subspaces at varying places. This methodology improves the ability to concentrate on various segments of the input sequence and obtain a more thorough comprehension of the connections within the data.

Operating the attention mechanism in parallel multiple times, each time employing a distinct learned linear projection of the queries, keys, and values, is the underlying concept of multi-head attention. This functionality enables the model to discern various forms of relationships within the data since every "head" can concentrate on distinct characteristics and facets of the input sequence. The MHSA is calculated in the following steps.

1. Linear Projections. The queries (Q), keys (K), and values (V) are linearly projected multiple times with different, learnable weight matrices, which can be presented by (2)

$$\begin{aligned} Q &= QW_i^Q, \\ K &= KW_i^K, \\ V &= VW_i^V \end{aligned} \tag{2}$$

   where W_i^Q,W_i^K,W_i^V are the weight matrices for the i_th head's linear transformations of Q, K, and V.

2. Scale Dot-Product Attention. Each head computes attention on its respective projections, using a scaled dot-product attention mechanism. This involves calculating dot products of the queries with all keys, dividing each by $\sqrt{d_k}$, and applying a softmax function to obtain weights on the values. The Scale Dot-Product Attention can be expressed as (3):

$$\begin{aligned} &\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ &\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{3}$$

3. Concatenation and Final Linear Projection. The outputs from each head are concatenated and then linearly transformed into the expected dimensions to acquire the final output of the MHSA.

$$\text{MHSA}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{4}$$

where Concat is the concatenation operation and W^O denotes the weight matrix for the final linear transformation.

2.3. Point-Wise Feed-Forward Network

In the context of a Transformer, the feed-forward network (FFN) is uniformly and independently applied to each location. This implies that every location in the output of the encoder or decoder, which refers to the representation of each word or token, undergoes the same FFN. However, the FFN functions independently at each position. A typical configuration of an FNN is the inclusion of two linear transformations, separated by a Rectified Linear Unit (ReLU) activation function. The first linear transformation maps the input onto a space with a higher number of dimensions (denoted as $d_f f$), whereas the subsequent linear transformation maps it back to the original lower-dimensional space of the model (denoted as d_{model}).

$$\text{FFN} = Relu(\text{MHSA} \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{5}$$

where MHSA is the output of the previous of MHSA block; *Relu* denotes the is the rectified linear activation function; and W_1, W_2, b_1, b_2 are trainable parameters of the FFN.

3. The Methodology for EGT Prediction

3.1. Overall Architecture of the Proposed Method

The overall architecture of the proposed Enhanced Scale-Aware Efficient Transformer (ESAE-Transformer) is depicted in Figure 2. The suggested method primarily comprises Mix data Embedding, the Multi-Head ProbSparse Attention (MHPSA)-based Encoder, the Multi-Scale Feature Aggregation Module (MSFAM), and the hybrid attention-based decoder. The proposed model takes the unprocessed sensor data as input and generates a prediction for the EGT as output. The Mix data Embedding is specifically utilized before the encoder and decoder sections to transform the raw sensor data into a consistent dimension, enabling the acquisition of positional information and learnable features. The encoder component consists of a sequence of Multi-Head ProbSparse Attention and point-wise feed-forward neural networks. Its primary function is to encode the embedded data into a nonlinear space with high dimensions. The proposed Multiscale Feature Aggregation Module (MSFAM) is utilized to extract significant temporal characteristics across several scales in a high-dimensional domain. The decoder component consists of two attention modules and two feed-forward neural network (FFN) blocks. The attention module consists of two components: the fundamental Multi-Head Self Attention and the enhanced Multi-Head ProbSparse Attention. Ultimately, we obtained the anticipated EGT from a fully-connected output layer. The intricate configurations of the indicated modules are presented in the subsequent sections.The specific parameters settings of proposed model are shown in Table 1, where N represents the number of samples, L represents the sequence length of time sliding window, and D denotes the dimension of the model that will be discussed in Section 5.1.

Table 1. The detailed input and output of each module.

Module	Block	Input Size	Output Size
Mix data Embedding	Position embedding	$(N, L, 8)$	(N, L, D)
	Learnable embedding	$(N, L, 8)$	(N, L, D)
Encoder	MHPSA	(N, L, D)	(N, L, D)
	FFN	(N, L, D)	(N, L, D)
MSFAM		(N, L, D)	(N, L, D)
Decoder	MHPSA	(N, L, D)	(N, L, D)
	FFN	(N, L, D)	(N, L, D)
	MHSA	(N, L, D)	(N, L, D)
	FFN	(N, L, D)	(N, L, D)
FC Layer		(N, L, D)	$(N, 2, 1)$

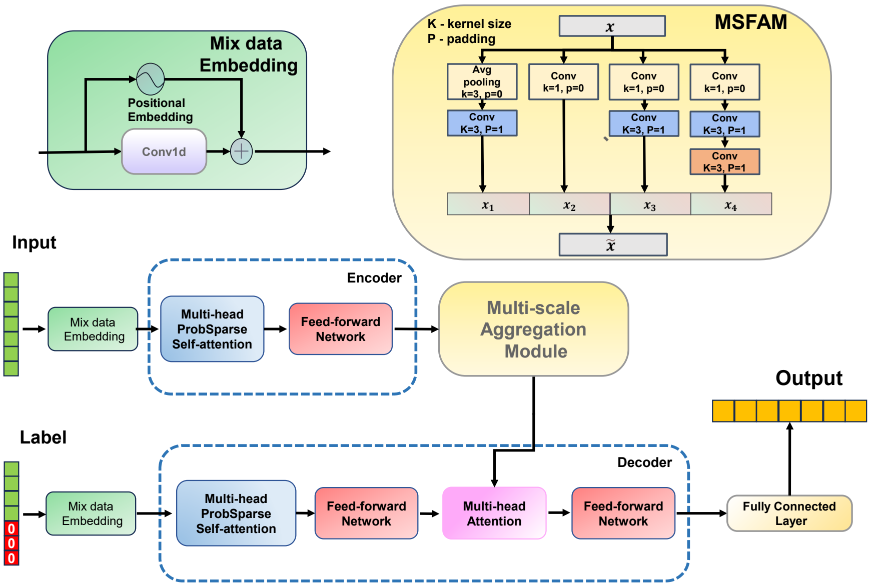


Figure 2. Theoverall structure of the proposed multi-scale enhanced efficient Prob-Transformer.

3.2. Mix Data Embedding

In the context of time series prediction tasks, it is important to address the limitations of positional token embedding. To overcome these limitations, the convolutional token embedding technique is employed to integrate these augmented features [25]. We designed a learnable convolutional token embedding PC, which can be expressed as (6):

$$PC = LRelu(\mathcal{C}(x_i)), k = 3, p = 1 \tag{6}$$

where $LRelu$ denotes the Leaky Relu activation function; $\mathcal{C}$ denotes the convolution layer; and k and p represents the kernel size and padding, respectively. We designed a mix token embedding by adding the position embedding PE in the (1) with the PC together, which is represented by (7):

$$P_H = PE + PC \tag{7}$$

3.3. Multi-Head ProbSparse Self-Attention Enhanced Encoder

In the context of enhancing the efficiency of neural network architectures for handling, particularly in time-series forecasting, the ProbSparse Self-Attention mechanism, as incorporated in the Informer model, emerges as a noteworthy innovation [26]. This mechanism is distinctively designed to address the computational constraints of traditional self-attention mechanisms. It achieves this by concentrating each key's attention on a subset of queries, specifically the most dominant ones within the sequence, thereby significantly reducing the computational load.

Crucially, the ProbSparse Self-Attention mechanism employs a sparse query matrix, denoted as $\overline{Q}$, which is a streamlined version of the original query matrix Q. This sparse matrix is composed exclusively of the Top-u queries, selected based on a sparsity measurement denoted as $M(q, K)$ [26]. The number of dominant queries, u, is determined by a constant sampling factor, c, and is calculated as $c \cdot \ln L_Q$, where L_Q signifies the length of the query sequence. This strategic selection process ensures that the computational complexity for each query–key interaction is substantially reduced to $O(\ln L_Q)$, marking a significant departure from the quadratic complexity observed in standard self-attention mechanisms.

The mathematical framework of the ProbSparse Self-Attention mechanism is articulated through the Equation (8)

$$\text{ProbSparse Self-Attention}(Q, K, V) = Softmax\left(\frac{\overline{Q}K^T}{\sqrt{d_k}}\right)V \tag{8}$$

where the sparse query matrix $\overline{Q}$ encapsulates the selected Top-u queries. This formulation effectively balances computational efficiency and the richness of the attention mechanism.

From a practical standpoint, the ProbSparse Self-Attention mechanism optimizes memory usage, maintaining it at $O(L_K \ln L_Q)$, where L_K is the length of the key sequence. This optimization is particularly advantageous compared to conventional self-attention frameworks. Moreover, in multi-head attention configurations, this mechanism ensures the generation of diverse sparse query–key pairs for each head, thereby mitigating potential information loss. Furthermore, to address the computational demands of determining the sparsity measurement for all queries, an empirical approximation approach is proposed. This approach precludes the necessity for exhaustive $O(L_Q L_K)$ dot-product computations, thereby enhancing efficiency. Additionally, the mechanism is designed to counter potential numerical stability issues, specifically in operations such as LogSumExp (LSE).

Overall, the ProbSparse Self-Attention mechanism represents a significant advancement in the design of neural network models for processing long sequential data, striking a notable balance between computational efficiency and the effective utilization of the attention mechanism. This makes it particularly suitable for applications that involve extensive time-series prediction analysis.

3.4. Multi-Scale Feature Aggregation Module

Inspired by the work in [27,28], we meticulously developed the Multi-Scale Feature Aggregation Module to explore the encoded high-dimensional feature in more detail. This module enhances the model's receptive field and uncovers implicit information from many scales inside the nonlinear feature space. The details of the MSFAM are shown in Figure 2. The MSFAM has four branches b_k with $k = 1, 2, \ldots, 4$. The calculation of each branch can be expressed as (9):

$$\begin{cases} b_1 = Conv(Avgpool(x))k_1 = 3, p_1 = 0, k_2 = 3, p_2 = 1 \\ b_2 = Conv_1(x), k_1 = 1, p_1 = 0 \\ b_3 = Conv_2(Conv_1(x)), k_1 = 1, p_1 = 0, k_2 = 3, k_3 = 3 \\ b_4 = Conv_3(Conv_2(Conv_1(x))), k_1 = 1, p_1 = 0, k_2 = 3, p_2 - 1, k_3 = 3, p_3 = 1 \end{cases} \tag{9}$$

where $Conv$ represents 1-D convolution, $Abgpool$ represents the average pooling layer, k denotes the kernel size of the convolution layer, and p denotes the padding size of the convolution layer. The output of last four branches is concatenated to acquire the multi-scale fused feature $\hat{x}$. The specific input and output of each branch is shown in the Table 2.

Table 2. The detailed input and output in each branch of MSFAM.

Branch	Layer	Input Size	Output Size
b_1	Avg pooling	(N, D, L)	(N, D, L)
	Conv_1	(N, D, L)	$(N, \frac{D}{4}, L)$
b_2	Conv_1	(N, D, L)	$(N, \frac{D}{4}, L)$
b_3	Conv_1	(N, D, L)	$(N, \frac{D}{6}, L)$
	Conv_2	$(N, \frac{D}{6}, L)$	$(N, \frac{D}{4}, L)$
b_4	Conv_1	(N, D, L)	$(N, \frac{D}{6}, L)$
	Conv_2	$(N, \frac{D}{6}, L)$	$(N, \frac{D}{4}, L)$
	Conv_3	$(N, \frac{D}{4}, L)$	$(N, \frac{D}{4}, L)$

3.5. Series-Attention-Based Decoder

The decoder component of our proposed architecture encompasses two attention blocks and two feed-forward networks (FFNs). This module's input includes the embedded label and the output-enhanced features produced by the Multi-Scale Attention Feature Module (MSFAM). As depicted in Figure 2, the label data are first processed through the masked Multi-Head ProbSparse Self-Attention (MHPSA) module. This module's primary objective is to discern the dependencies within the label data. Subsequently, the Multi-Head Self-Attention (MHSA) mechanism focuses on learning the correlations between the enhanced high-level features across diverse time steps, effectively integrating the outputs of the sequentially arranged MHPSA and FFN.

During the training phase, the Exhaust Gas Temperature (EGT) label data are meticulously prepared, allowing the decoder to concurrently decode each time step, leveraging the model's inherent parallel computing capabilities. This process ensures that the information pertaining to future label data is comprehensively learned at each time step during the MHPSA computation. This learned information is then incorporated into the MHSA for further processing. Consequently, to prevent the unintended leakage of future label data, a masking operation is crucial within the MHPSA module of the decoder. This necessitates the introduction of a specific matrix in the scaled dot-product attention mechanism. The matrix, characterized by a lower triangle and diagonal populated with 1 s and an upper triangle filled with 0 s, is instrumental in obfuscating future information, thus maintaining the integrity of the predictive process.

4. Experiment Setting

4.1. Data Description

This study utilizes a dataset derived from the Quick Access Recorder (QAR) data, obtained from a commercial aircraft, to evaluate the proposed methodology. QAR data, pivotal in modern aviation, encompasses digital recordings of various flight characteristics and system information. These recordings are gathered from an aircraft's sensors and systems during flight operations. The data collection is facilitated by the Quick Access Recorder (QAR), a device installed on the aircraft. QAR data fulfill several key roles, extending beyond maintenance and safety analysis. They are instrumental in monitoring flight performance and investigating incidents. Their significance in the aviation industry is multifaceted as they provide essential insights for airlines, maintenance teams, and regulatory bodies. These data aid in overseeing and assessing the operational efficiency and safety of aircraft. One of the primary advantages of using QAR data is their potential in the early detection and intervention of emerging issues, significantly reducing the risk of these concerns evolving into more severe problems. Furthermore, the application of QAR data can lead to improvements in operational efficiency, underscoring their value in both preventive measures and enhancing overall aviation operations.

This research employs a dataset comprising aircraft engine data, recorded across various phases of each flight mission, including climb, cruise, and landing. The dataset encompasses a comprehensive account of the engine's operational processes, capturing both transient and steady-state phases. Transient state prediction presents a more challenging task than steady-state process prediction due to the dynamic nature of the former. The dataset, sourced from the Quick Access Recorder (QAR), documents over 200 distinct parameters, with a sampling frequency of 4 Hz for condition parameters. For the purposes of this study, the data underwent downsampling to a 1 Hz sample rate, aligning with the universal QAR data standards.

A careful selection of parameters was made from the extensive pool of over 200, focusing on those most indicative of gas path performance in the engine, as referenced in prior studies [13,21,29]. Scenario descriptors, including ALT, MN, PLA, and T0, are crucial in determining the flying state and are imperative for accurate Exhaust Gas Temperature (EGT) prediction. Taking a dual shaft engine as an example, the sensor positions used to measure the above parameters are shown in the Figure 3. Additionally, incorporating gas path data such as rotational speeds, temperatures, or pressures could further refine the precision of predictive models.

The parameters selected for this study are detailed in Table 3. The initial seven parameters listed are believed to have a direct correlation with EGT. In contrast, the final parameter, EGT itself, is the focal point of prediction. Figure 4 presents a visual representation of the raw sensor data across a complete flight cycle, providing an empirical basis for the analysis conducted in this study. The dataset utilized in this study captures data from the aircraft engine during each flying mission's climb, cruise, and landing phases. The continuous process of the engine in operation, including the transient and steady-state processes, is contained in this dataset. Predicting transient states is more difficult than steady-state process prediction. More than 200 distinct parameters are recorded in this QAR dataset, with a 4 Hz sample frequency for the condition parameters. To establish the model, we carry out downsampling at a sample rate of 1Hz for universal QAR data.

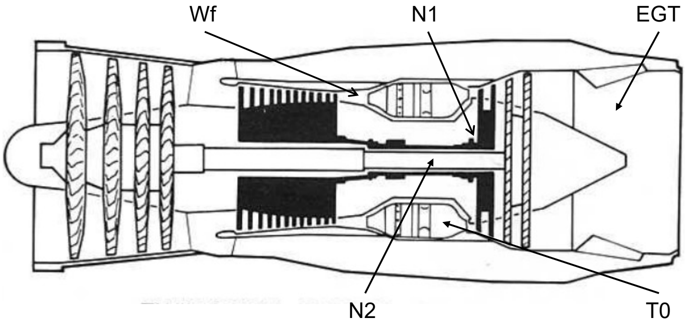


Figure 3. Schematic diagram of engine sensor location.

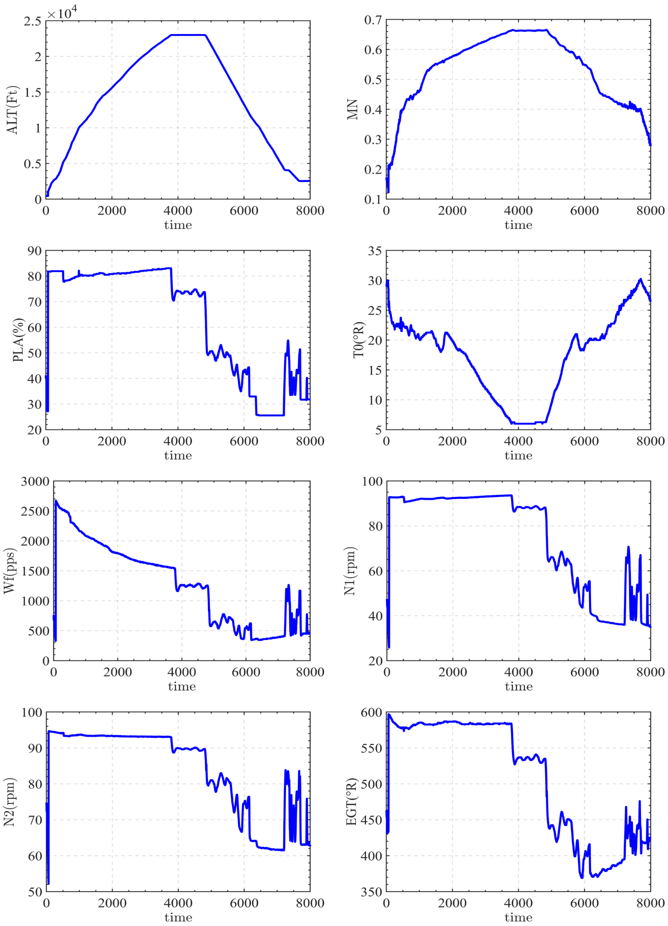
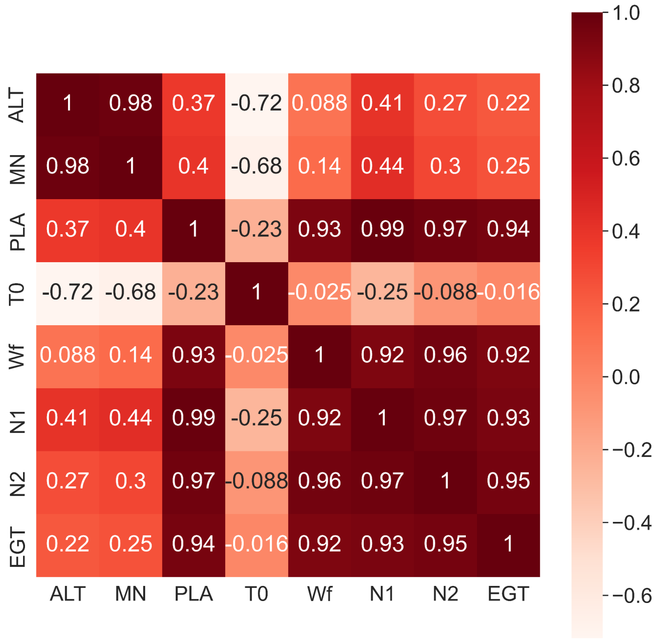


Figure 4. The raw sensor data description of a flight cycle.

Table 3. The description of the dataset.

No. of Parameters	Symbols	Description	Units
1	ALT	Flight altitude	Ft
2	MN	Mach Number	-
3	PLA	Power lever angle	%
4	T0	Ambient temperature	$^{\circ}R$
5	Wf	Fuel flow rate	pps
6	N1	Physical fan speed	ppm
7	N2	Physical core speed	ppm
8	EGT	Exhaust gas temperature	$^{\circ}R$

To provide a more thorough investigation of the raw sensor data with the aim of constructing a predictive model, the implementation of correlation analysis is employed to evaluate the temporal link among the sensor data. In this analysis, we utilize the Spearman correlation coefficient (SCC) to examine the data, which quantifies the correlation using a monotonic function. It is evident that the Spearman correlation coefficient possesses the capability to handle data that contain outliers, non-normal distributions, or heteroscedasticity [30]. The result of the SCC of the selected parameters is indicated in Figure 5.



	ALT	MN	PLA	T0	Wf	N1	N2	EGT
ALT	1	0.98	0.37	-0.72	0.088	0.41	0.27	0.22
MN	0.98	1	0.4	-0.68	0.14	0.44	0.3	0.25
PLA	0.37	0.4	1	-0.23	0.93	0.99	0.97	0.94
T0	-0.72	-0.68	-0.23	1	-0.025	-0.25	-0.088	-0.016
Wf	0.088	0.14	0.93	-0.025	1	0.92	0.96	0.92
N1	0.41	0.44	0.99	-0.25	0.92	1	0.97	0.93
N2	0.27	0.3	0.97	-0.088	0.96	0.97	1	0.95
EGT	0.22	0.25	0.94	-0.016	0.92	0.93	0.95	1

Figure 5. The Spearman correlation cofficient matrix of the selected parameters.

4.2. Data Preparation

4.2.1. Time Sliding Processing

The exploration of associations among adjacent data points from different time intervals is paramount in time series data analysis [31]. This research utilizes the sliding window technique, which entails segmenting data points across various time intervals using designated time windows. This method is crucial for capturing the temporal relationships inherent in the data. To augment the sample size, the sliding stride is set to a value of 1. Figure 6 illustrates this sliding window process. At each time point, the input features correspond to the Exhaust Gas Temperature (EGT) values. In practical scenarios, the selection of the appropriate window size is determined based on the characteristics of the raw data, ensuring that the analysis is grounded in real-world application contexts.

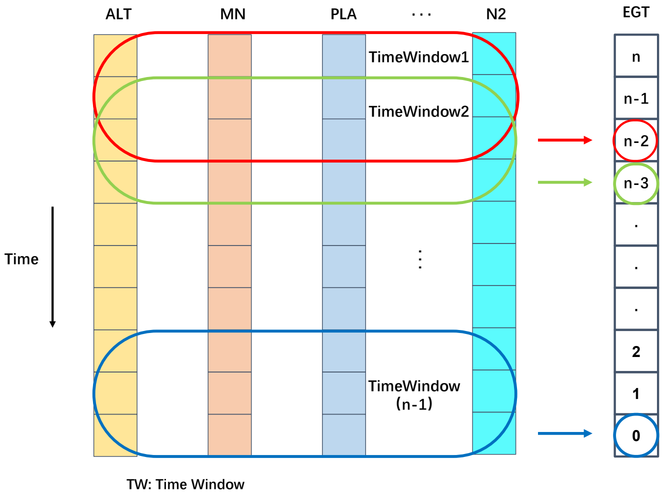


Figure 6. Example of time sliding window data segmentation for the EGT prediction.

4.2.2. Data Normalization

The mean-standard deviation is a statistical measure that quantifies the dispersion or variability of a dataset around its mean. It is calculated by normalization, sometimes referred to as Z-Score normalization or standardization, is a widely employed data preparation method utilized in the fields of statistics and machine learning. The process entails transforming the characteristics of the dataset in order to conform to a typical normal distribution, characterized by a mean of 0 and a standard deviation of 1. This approach proves to be particularly advantageous in situations when it is necessary to normalize the data without compromising the integrity of the variations in the range of values [32]. The formula for Z-Score normalization is expressed in (10):

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{10}$$

where x_{norm} is the normalized values; x denotes the original data; μ and σ represent the mean and standard deviation of the data, respectively.

4.3. Evaluation Metrics

Regarding the EGT outcomes, two commonly employed metrics are utilized to evaluate the effectiveness of the presented models. The first metric is the Root Mean Square Error (RMSE), while the second metric is referred to as the Mean Absolute Error (MAE).

1. RMSE. Root Mean Square Error (RMSE) is a commonly used metric for quantifying the disparities between the projected values generated by a model and the actual observed values. The use of this approach becomes advantageous in situations when the potential impact of significant errors surpasses that of minor errors since it assigns a substantially greater significance to the occurrence of big errors [33]. The RMSE can be expressed by (11):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

   where n represents the number of ground truth values, while y_i and $\hat{y}_i$ represent the predicted and actual values, respectively.
2. MAE. The Mean Absolute Error (MAE) quantifies the average size of mistakes within a given collection of predictions, disregarding their directional aspect. The academic formulation involves calculating the mean absolute difference between the predicted

values and the actual observations in a test sample, with each individual difference being assigned equal weight [33]. The MAE is implemented by (12):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

In the given context, the variable n denotes the count of ground truth values, while y_i and $\hat{y}_i$ represent the predicted and actual values correspondingly.

4.4. Implement Details

The research was conducted via an electronic platform that consisted of a workstation outfitted with various hardware and software configurations. The workstation is equipped with an Intel Core i7-10870H central processing unit (CPU) that operates at a clock speed of 2.20 gigahertz (GHz). In addition, the machine is equipped with a high-performance NVIDIA Geforce RTX3060 graphics processing unit (GPU), hence enhancing the computational capabilities necessary for carrying out the study. The workstation is furnished with a significant amount of RAM, namely, 64GB, which ensures enough memory capacity for data processing and model training tasks. The suggested model was subjected to training and optimization via the AdaX optimization approach. The AdaX method, as introduced by [34], is an innovative strategy for optimizing adaptive learning rates. The algorithm in question integrates the favorable characteristics of the AdaGrad and Adam algorithms, resulting in enhanced convergence and generalization abilities.The training operation started with an initial learning rate of 0.001, which is scheduled to undergo ten decays following the completion of 10 epochs.

5. Results and Discussion

In our rigorous evaluation of the proposed method for practical application, we conducted several experiments under diverse conditions. These experiments were specifically designed to analyze the impact of varying prediction lengths and input sequence settings on the method's performance. The prediction length was fixed at 2 s, while the input sequence length was varied, including settings of 16 s, 8 s, 4 s, and 2 s. This comprehensive approach allowed for a detailed assessment of how different input sequence lengths influence the model's predictive accuracy and efficiency. Such an analysis is critical in understanding the model's adaptability and effectiveness across a range of temporal scales, providing valuable insights for its application in real-world scenarios.

5.1. The Impact of the Feature Dimension of the Model

In the aforementioned tables, denoted as Tables 1 and 2, a comprehensive illustration of the encoder, decoder, and (MSFAGM) is presented, highlighting their shared feature dimension, represented as D. The significant role played by the feature dimension D in influencing the predictive outcomes is evident from these representations as shown in Table 1. It modifies the dimension of feature space in the model, which is the model dimension parameter. The schematic of the model dimension is illustrated in the Figure 7. This crucial aspect prompted an in-depth exploration, where the proposed method was subjected to rigorous training across a spectrum of varying feature dimensions. Such an investigation is pivotal in discerning the optimal feature dimension that maximizes the efficacy of the predictive model, thereby ensuring enhanced performance in practical applications. This study underscores the intricate relationship between the feature dimension and the model's predictive accuracy, providing valuable insights for further optimizations.

Firstly, we will introduce the experimental process in Table 4. The output of the model is a fixed time series of 2 s in length, while the input of the model is an indefinite time series, and the length of the input is artificially determined. As shown in Figure 8, the area marked by two red dashed lines is the result that the model needs to predict, and it displays the predicted value of EGT in 16–18 s. To maximize the performance of the model, we conducted experiments on the model from two aspects. On the one hand, it is to change the input length of the model. We use the time series of the first 2 s, 4 s, 8 s, and 16 s of the predicted sequence as model inputs.

Table 4. The result under the different dimensions of the proposed method.

No.	Model Dimension	16 s		8 s		4 s		2 s	
		RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
No. 1	32	9.51	6.83	8.09	5.39	8.20	5.67	15.29	8.54
No. 2	64	8.91	6.11	7.17	4.76	7.12	4.79	6.94	4.55
No. 3	128	**6.69**	**4.51**	**6.45**	4.20	**6.32**	**3.84**	**5.94**	**3.47**
No. 4	256	7.03	4.52	6.74	**4.06**	6.48	3.86	9.71	4.45
No. 5	512	13.65	9.98	12.67	9.34	9.36	6.67	9.18	5.52

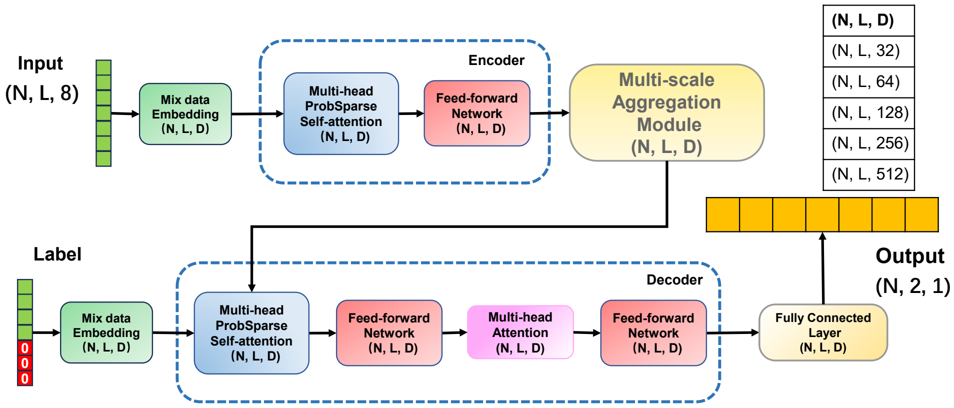

Input
(N, L, 8)
Mix data Embedding (N, L, D)
Encoder
Multi-head ProbSparse Self-attention (N, L, D)
Feed-forward Network (N, L, D)
Multi-scale Aggregation Module (N, L, D)
(N, L, D)
(N, L, 32)
(N, L, 64)
(N, L, 128)
(N, L, 256)
(N, L, 512)
Label
0
0
0
Mix data Embedding (N, L, D)
Decoder
Multi-head ProbSparse Self-attention (N, L, D)
Feed-forward Network (N, L, D)
Multi-head Attention (N, L, D)
Feed-forward Network (N, L, D)
Fully Connected Layer (N, L, D)
Output
(N, 2, 1)

Figure 7. The input and output description of each module.

The observations drawn from Table 4 reveal a nuanced trend in the relationship between the model dimension and its overall prediction accuracy. It is intuitively apparent that as the model dimension increases, there is an initial uptrend in prediction accuracy, which is subsequently followed by a downtrend. Notably, the model achieves its lowest error metrics when the dimension value is set to 128. This phenomenon is consistently observed across various time points.

Additionally, a horizontal comparison of the prediction effects at different prediction times elucidates that the highest overall prediction accuracy is attained when the prediction time is set to 2 s. Based on these empirical findings, the paper strategically adjusts the feature dimension to 128 for subsequent method comparisons. This adjustment is premised on optimizing the model's performance, ensuring that it is attuned to deliver the highest prediction accuracy under these specific parameters. Figure 9 presents a random selection of prediction results, randomly chosen to illustrate performance across varying input time lengths and feature dimensions. It is discernible that the input time length of 2 s consistently yields the most accurate predictions. Consequently, for a more granular analysis, we have delineated the 2 s input length prediction results in Figure 10, categorizing them into distinct periods of rise, decline, and fluctuation. This segmentation facilitates a deeper understanding of the model's predictive behavior under dynamic conditions, highlighting its performance nuances in response to temporal variations.

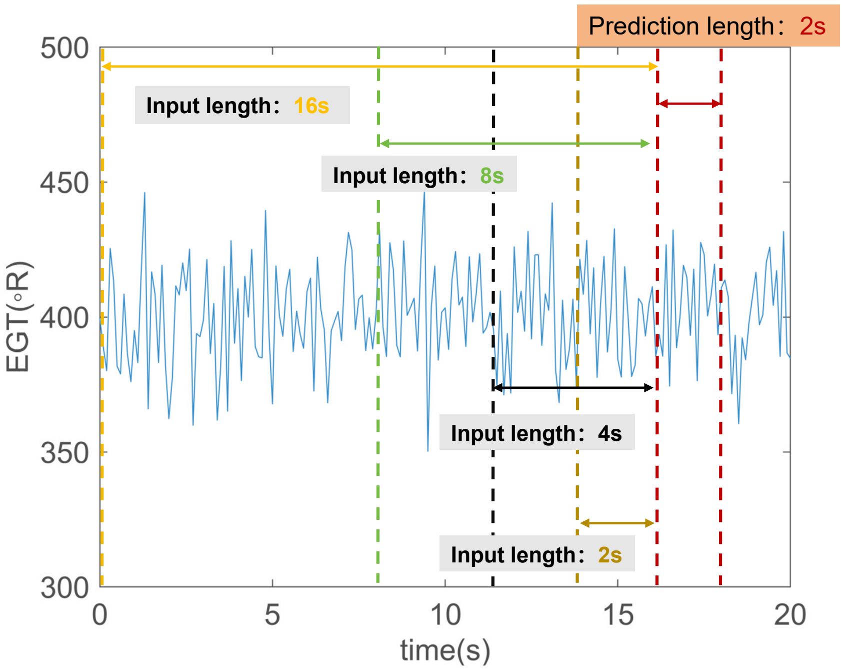


Figure 8. The description of input length and output length for the proposed model.

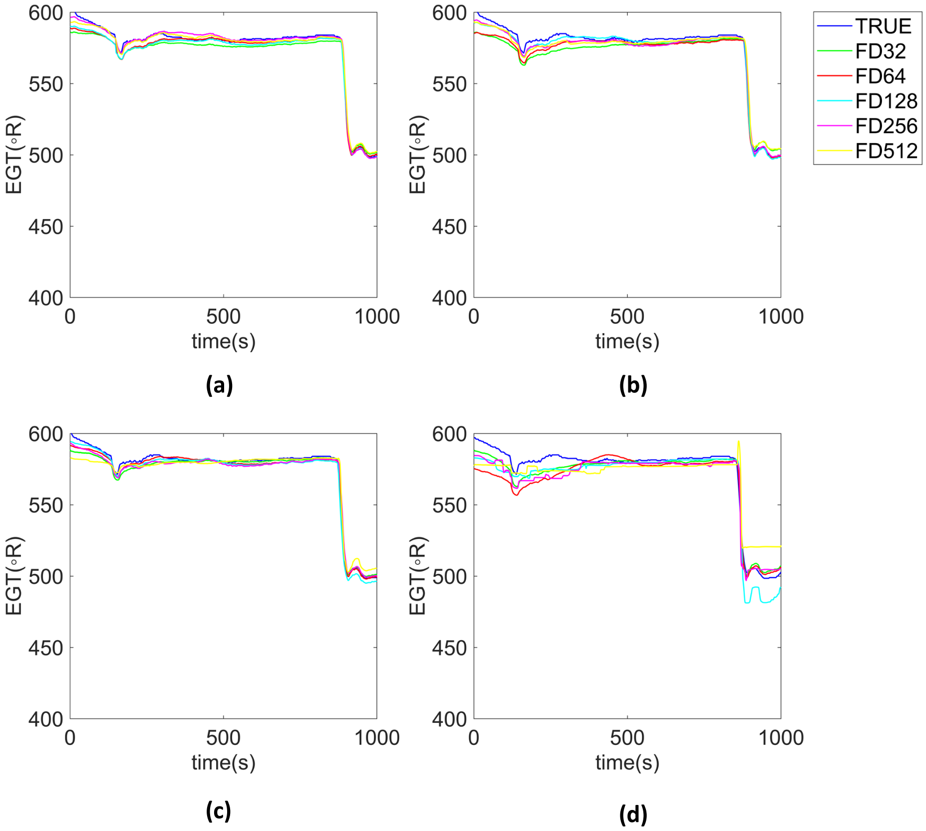


Figure 9. Randomly selected EGT prediction results: a comparative analysis across diverse input time lengths and feature dimensions. (**a**) 2 s; (**b**) 4 s; (**c**) 8 s; and (**d**) 16 s.

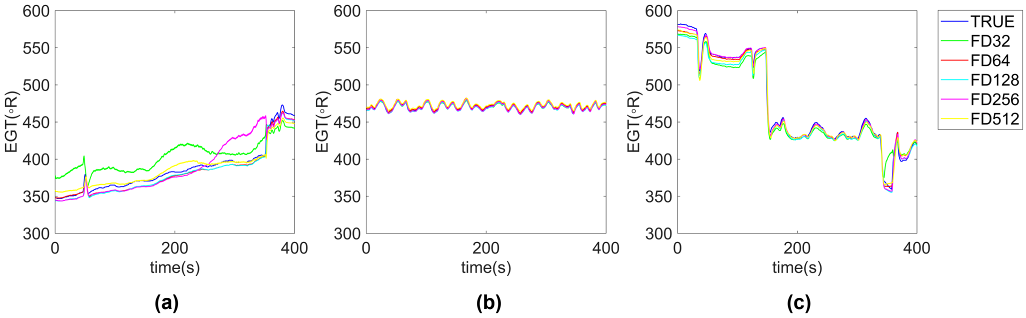


Figure 10. Randomly selected EGT prediction results: a comparative analysis across diverse feature dimensions under different phases. (**a**) rising; (**b**) fluctuating; and (**c**) decreasing.

5.2. Compared with Other State-of-Art Methods

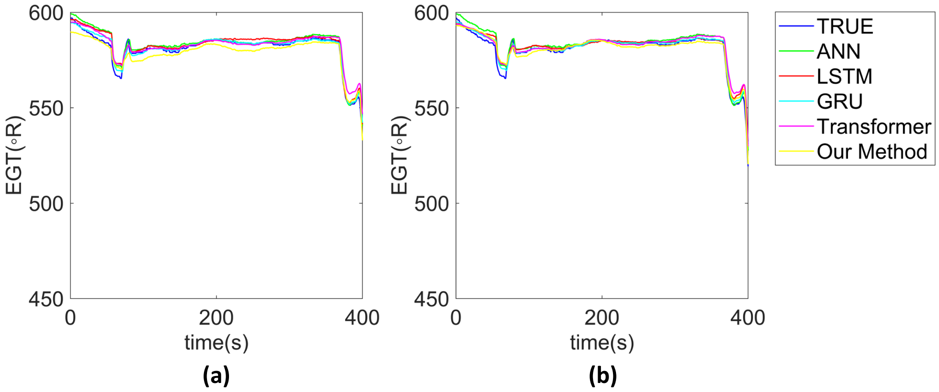
Our model exhibits a significant advantage in forecasting accuracy when compared to other contemporary models. To provide a comprehensive and rigorous comparison, we evaluated our proposed model against four state-of-the-art EGT prediction models, each representing a unique approach within the domain. These include the Artificial Neural Network (ANN), long short-term memory (LSTM), Gated Recurrent Unit (GRU), and Transformer models. This comparative analysis was conducted under varying dimensions to ensure a thorough assessment of each model's capabilities. The decision to include these specific models stems from their widespread recognition and established efficacy in EGT forecasting. The ANN serves as a foundational model in neural network research, offering a baseline for comparison. The LSTM and GRU, both of which are variants of recurrent neural networks, are renowned for their ability to capture long-term dependencies in sequential data, a crucial aspect in accurate time series forecasting. The Transformer, known for its self-attention mechanism, represents the cutting edge in handling sequential data and offers a contrast to the recurrent architectures of LSTM and GRU. By comparing our proposed model with these diverse and well-regarded models, we aim to demonstrate its superior forecasting accuracy across various dimensions. This comparison not only underscores the strengths of our model but also contributes to a deeper understanding of its performance in the broader context of EGT prediction methodologies.

Table 5 effectively highlights the superior predictive performance of our proposed methods, as evaluated by two key metrics. This enhancement is particularly striking when contrasted with various baseline models. In such comparisons, our method demonstrates exceptional testing performance, especially in scenarios characterized by shorter prediction times. Figure 11 presents a selection of predictions with varying input lengths, illustrating a consistent trend where shorter input sequences result in improved performance. Notably, with an input length of 2 s, the prediction's Mean Absolute Error (MAE) shows a significant improvement of 6.9%, and the Root Mean Square Error (RMSE) registers a 3.4% enhancement when compared to the standard Transformer model. This superior performance extends beyond the Transformer model, also surpassing the results of ANN, LSTM, and GRU models. This clearly indicates the proficiency of the self-attention-based structure in effectively capturing temporal features.

Practically, a shorter input sequence length implies reduced computational load. For instance, with a 2 s input, the Mean Absolute Error prediction of $3.47^{\circ}R$ aligns suitably with real-world environmental conditions. Additionally, Figure 12 delineates the rising, fluctuating, and decreasing phases of the EGT prediction. This dichotomy in performance can be attributed to the Transformer's inherent capability to adeptly capture short-term local semantic interactions, making it an ideal choice for time series prediction tasks.

Table 5. The prediction performance of all the models under different prediction times.

Method	16 s		8 s		4 s		2 s	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ANN	8.89	5.14	8.07	4.97	8.36	4.73	8.03	4.34
LSTM	8.91	6.18	7.17	4.76	7.12	4.79	6.94	4.55
GRU	10.88	7.45	7.07	4.51	7.36	4.38	7.39	4.28
Transformer	9.94	7.06	6.77	4.24	6.80	4.26	6.15	3.73
The proposed method	**6.69**	**4.51**	**6.45**	**4.20**	**6.32**	**3.84**	**5.94**	**3.47**



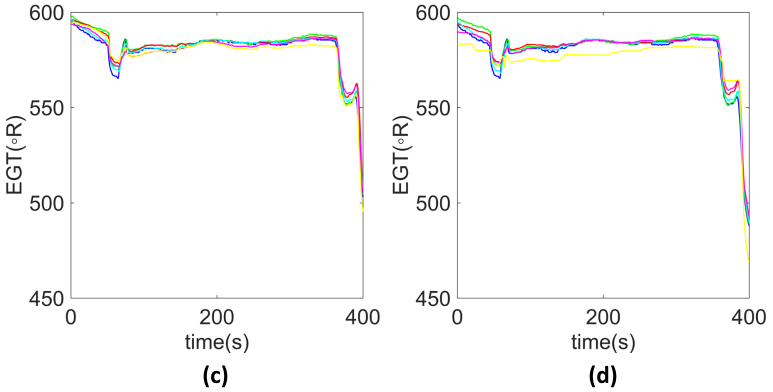


Figure 11. Randomly selected EGT prediction results: a comparative comparison with other State-of-Art methods across diverse input time lengths. (**a**) 2 s; (**b**) 4 s; (**c**) 8 s; and (**d**) 16 s.

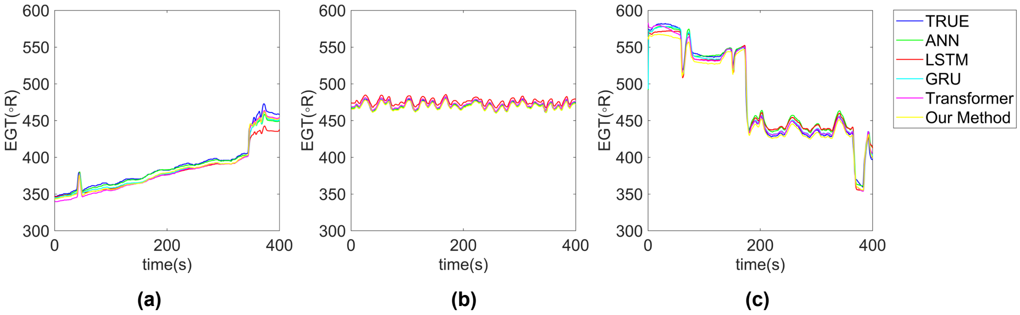


Figure 12. Randomly selected EGT prediction results: a comparative comparison with other State-of-Art methods across under different phases. (**a**) rising; (**b**) fluctuating; and (**c**) decreasing.

5.3. Ablation Study of the Proposed Method

To thoroughly evaluate the individual contributions of the various components within our proposed model, we embarked on an ablation study. This study was meticulously designed to dissect the impact of each component on the overall predictive efficacy of the model. To ensure a comprehensive analysis, the ablation study was conducted under a range of input sequence lengths, providing insights into how different components perform under varying temporal scales.

In this study, the Transformer model was adopted as the baseline. This choice is strategic as the Transformer's architecture, renowned for its self-attention mechanism, offers a robust foundation for comparison. By systematically removing or altering specific components of our proposed model and comparing the resultant performance against the baseline Transformer, we can isolate and understand the contribution of each individual component. The result is shown in Table 6.

Table 6. The results of different component ablation experiments with multiple input lengths.

No.	Model Structure	16 s		8 s		4 s		2 s	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
No. 1	Transformer	9.94	7.06	6.77	4.24	6.80	4.26	6.15	3.73
No. 2	Transformer encoder + MSFAM + Transformer decoder	8.06	5.41	7.04	4.29	6.88	3.84	7.26	4.27
No. 3	MHPSA encoder+MHPSA decoder	7.01	**4.44**	6.75	4.31	6.36	4.00	6.39	3.97
No. 4	Transformer encoder+MSFAM+MHPSA decoder	7.96	5.61	7.26	4.83	8.15	4.52	6.83	3.99
No. 5	MHPSA encoder + MSFAM + Transformer decoder	7.18	4.61	7.55	**4.07**	7.42	4.01	6.60	3.84
No. 6	MHPSA encoder + MSFAM + MHPSA decoder	6.69	4.51	**6.45**	4.20	**6.32**	**3.96**	**5.94**	**3.47**

1. Experiment No. 1: The Baseline Transformer Model. This initial experiment establishes a baseline by employing a standard Transformer model. It serves as a reference point for evaluating the enhancements achieved in subsequent experimental configurations.
2. Experiment No. 2: The Integration of MSFAM. This trial involves the incorporation of the Multi-Scale Feature Aggregation Module (MSFAM) into the Transformer framework. The primary aim is to investigate how MSFAM's inclusion affects the model's predictive capabilities. Notably, this integration leads to improved prediction accuracy, particularly with longer input sequences, when compared to the pure Transformer model.
3. Experiment No. 3: The implementation of MHPSA.In this configuration, the Multi-Head ProbSparse Attention (MHPSA) mechanism is integrated into both the encoder and decoder components of the model. The objective is to examine the overall impact of MHPSA on the model's performance. The introduction of MHPSA is observed to enhance prediction accuracy consistently across all input sequence lengths, with a marked improvement in the 16 s input, yielding the best Mean Absolute Error (MAE) of 4.44.
4. Experiment No. 4: MHPSA with the Transformer Encoder. This experiment evaluates the effectiveness of a model configuration that combines a standard Transformer encoder with the MSFAM and an MHPSA decoder. The results indicate that a pure MHPSA decoder is particularly beneficial for longer input sequences.
5. Experiment No. 5: MSPHA with the Transformer Decoder. This setup is a reversal of the previous experiment, featuring an MHPSA encoder, the MSFAM, and a standard Transformer decoder. The focus is to assess the impact of incorporating MHPSA in the encoder while maintaining the traditional Transformer decoder. The findings suggest that this configuration is advantageous for relatively long input sequences.
6. Experiment No. 6: The MHPSA Encoder and Decoder with MSFAM. The final experiment combines MHPSA in both the encoder and decoder segments, along with the MSFAM. This setup aims to explore the synergistic effects of these components within a unified model. The results demonstrate exceptional performance in capturing temporal features, particularly achieving the best results with 4 s and 2 s input sequence lengths.

Each experiment in this series incrementally builds upon the previous one, allowing for a detailed analysis of how each modification contributes to the overall performance of the model. This systematic approach enables a nuanced understanding of the strengths and limitations of each component within the model's architecture, guiding further refinements and optimizations for enhanced predictive accuracy in EGT prediction.

5.4. Discussion

In the context of this study, the ESAE-Transformer model was meticulously developed and evaluated for its efficacy in predicting EGT in commercial aircraft, utilizing data from QAR. The assessment of the model's performance, quantified through MAE and Root Mean Square Error RMSE, was central to our analysis. Our experimental approach was methodically designed to explore the optimal configuration of the model, with a particular focus on the dimensionality of the feature space and the variation in input sequence lengths, all while maintaining a fixed prediction length of 2 s. When juxtaposed with traditional predictive models such as ANN, GRU, LSTM, and the contemporary Transformer models, ESAE-Transformer demonstrated a markedly superior performance, achieving an MAE of $3.47^{\circ}R$. This outcome not only validates the robustness of our model in the realm of EGT prediction but also underscores the potential of advanced analytical methods in enhancing aeronautical applications. However, it is imperative to acknowledge the limitations encountered in this study. While the model shows promising results, there is a discernible scope for augmenting its prediction accuracy. Moreover, the computational efficiency of the model, particularly in the context of real-time onboard application, requires further optimization. These aspects present avenues for future research, where the focus would be on refining the model to achieve higher accuracy and computational efficiency, thereby making it more viable for real-time deployment in aircraft systems.

6. Conclusions

In this paper, we proposed the Enhanced Scale-Aware efficient Transformer (ESAE-Transformer), an innovative Transformer-based model tailored for Exhaust Gas Temperature (EGT) prediction. The main contribution of this paper can be summarized as follows:

(1) We developed an innovative transformer-based model for predicting aero-engine exhaust gas temperature (EGT), marking a first in its application for estimating EGT in aero-engines;

(2) We developed the Multi-Head ProbSparse Self-Attention (MHPSA) mechanism in the encoder and decoder models to efficiently reduce temporal complexity and optimize memory usage, focusing on the most informative data segments;

(3) This paper implemented a Multi-Scale Feature Aggregation Module (MSFAM) to enhance the processing of complex temporal features, thereby improving the model's predictive accuracy for nuanced temporal dynamics;

(4) We conducted comprehensive evaluations, demonstrating optimal performance with a 2 s input length and a 128-dimensional model. The results of this comprehensive analysis indicate that the mean absolute prediction error is $3.47^{\circ}R$, which is well-aligned with real-world environmental conditions, underscoring the model's practical applicability.

Our future research endeavors will focus on enhancing the accuracy of Exhaust Gas Temperature (EGT) prediction while employing knowledge distillation techniques to streamline the methodology for real-time applications. This advancement aims to develop a lightweight, efficient model suitable for on-board implementation, thereby bridging the gap between high-precision predictive analytics and practical real-time operational requirements.

Author Contributions: Conceptualization, writing–original draft preparation, S.L.; validation, data preprocessing, N.Z.; visualization, C.S.; writing—review and editing, G.C.; and supervision, project administration, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

ESAE-Transformer	Enhance Scale-Aware Efficient Transformer
MSFAM	Multi-Scale Feature Aggregation Module
EGT	Exhaust Gas Temperature
AR	autoregressive
APIMA	autoregressive integrated moving average
RF	random forest
KF	Kalman filters
GRNN	Generalized Regression Neural Network
RBF	Radial Basis Function
SVR	Support Vector Regression
LSTM	long short-term memory
NARX	Nonlinear Auto-Regressive model with Exogenous Inputs
ARMA	autoregressive moving average
FAE	feature attention mechanism-enhanced
MA	Moving Average
RUL	remaining useful life
MHPSA	Multi-Head ProbSparse Self-Attention
NLP	natural language processing
MHSA	Multi-Head Self-Attention
FFN	feed-forward network
ReLU	Rectified Linear Unit
LSE	LogSumExp
QAR	Quick Access Recorder
ALT	Flight altitude
MN	Mach Number
PLA	Power lever angle
Wf	Fuel flow rate
SCC	Spearman correlation coefficient
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
ANN	Artificial Neural Network
GRU	Gate Recurrent Unit

References

1. Balakrishnan, N.; Devasigamani, A.I.; Anupama, K.; Sharma, N. Aero-engine health monitoring with real flight data using whale optimization algorithm based artificial neural network technique. *Opt. Mem. Neural Netw.* **2021**, *30*, 80–96. [CrossRef]
2. Ren, L.H.; Ye, Z.F.; Zhao, Y.P. A modeling method for aero-engine by combining stochastic gradient descent with support vector regression. *Aerosp. Sci. Technol.* **2020**, *99*, 105775. [CrossRef]
3. Li, H.; Huang, H.Z.; Li, Y.F.; Zhou, J.; Mi, J. Physics of failure-based reliability prediction of turbine blades using multi-source information fusion. *Appl. Soft Comput.* **2018**, *72*, 624–635. [CrossRef]
4. Wang, Z.T.; Zhao, N.B.; Wang, W.Y.; Tang, R.; Li, S.Y. A fault diagnosis approach for gas turbine exhaust gas temperature based on fuzzy c-means clustering and support vector machine. *Math. Probl. Eng.* **2015**, *2015*, 240267. [CrossRef]
5. Tuzcu, H.; Sohret, Y.; Caliskan, H. Energy, environment and enviroeconomic analyses and assessments of the turbofan engine used in aviation industry. *Environ. Prog. Sustain. Energy* **2021**, *40*, e13547. [CrossRef]
6. Li, D.; Peng, J.; He, D. Aero-engine exhaust gas temperature prediction based on LightGBM optimized by improved bat algorithm. *Therm. Sci.* **2021**, *25*, 845–858. [CrossRef]
7. Yildirim, M.T.; Kurt, B. Aircraft gas turbine engine health monitoring system by real flight data. *Int. J. Aerosp. Eng.* **2018**, *2018*, 9570873. [CrossRef]
8. Rao, B.N. The Role of Artificial Intelligence (AI) in Condition Monitoring and Diagnostic Engineering Management (COMADEM): A Literature Survey. *Am. J. Artif. Intell.* **2021**, *5*, 17–37.

9. Jiang, Y.; Yin, S.; Kaynak, O. Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond. *IEEE Access* **2018**, *6*, 47374–47384. [CrossRef]
10. Zhang, J.; Jiang, Y.; Wu, S.; Li, X.; Luo, H.; Yin, S. Prediction of remaining useful life based on bidirectional gated recurrent unit with temporal self-attention mechanism. *Reliab. Eng. Syst. Saf.* **2022**, *221*, 108297. [CrossRef]
11. Emer, N.; Özbek, N. A survey on Kalman filtering for unmanned aerial vehicles: Recent trends, applications, and challenges. In Proceedings of the International Conference on Engineering Technologies (ICENTE'20), Konya, Turkey, 19–21 November 2020; pp. 19–21.
12. Chen, Y.Z.; Li, Y.G.; Tsoutsanis, E.; Newby, M.; Zhao, X.D. Techno-economic evaluation and optimization of CCGT power Plant: A multi-criteria decision support system. *Energy Convers. Manag.* **2021**, *237*, 114107. [CrossRef]
13. Wang, Z.; Zhao, Y. Data-Driven Exhaust Gas Temperature Baseline Predictions for Aeroengine Based on Machine Learning Algorithms. *Aerospace* **2023**, *10*, 17. [CrossRef]
14. Li, H.Z.; Guo, S.; Li, C.J.; Sun, J.Q. A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowl.-Based Syst.* **2013**, *37*, 378–387. [CrossRef]
15. Aljarah, I.; Faris, H.; Mirjalili, S.; Al-Madi, N. Training radial basis function networks using biogeography-based optimizer. *Neural Comput. Appl.* **2018**, *29*, 529–553. [CrossRef]
16. Singh, R.; Arora, H.C.; Bahrami, A.; Kumar, A.; Kapoor, N.R.; Kumar, K.; Rai, H.S. Enhancing sustainability of corroded RC structures: Estimating steel-to-concrete bond strength with ANN and SVM algorithms. *Materials* **2022**, *15*, 8295. [CrossRef] [PubMed]
17. Gupta, K.K.; Kalita, K.; Ghadai, R.K.; Ramachandran, M.; Gao, X.Z. Machine learning-based predictive modelling of biodiesel production—A comparative perspective. *Energies* **2021**, *14*, 1122. [CrossRef]
18. Ullah, S.; Li, S.; Khan, K.; Khan, S.; Khan, I.; Eldin, S.M. An Investigation of Exhaust Gas Temperature of Aircraft Engine Using LSTM. *IEEE Access* **2023**, *11*, 5168–5177. [CrossRef]
19. Asgari, H.; Chen, X.; Morini, M.; Pinelli, M.; Sainudiin, R.; Spina, P.R.; Venturini, M. NARX models for simulation of the start-up operation of a single-shaft gas turbine. *Appl. Therm. Eng.* **2016**, *93*, 368–376. [CrossRef]
20. Pham, H.T.; Yang, B.S. A hybrid of nonlinear autoregressive model with exogenous input and autoregressive moving average model for long-term machine state forecasting. *Expert Syst. Appl.* **2010**, *37*, 3310–3317. [CrossRef]
21. Ma, S.; Wu, Y.; Zheng, H.; Gou, L. A Hybrid of NARX and Moving Average Structures for Exhaust Gas Temperature Prediction of Gas Turbine Engines. *Aerospace* **2023**, *10*, 496. [CrossRef]
22. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
24. Liu, L.; Song, X.; Zhou, Z. Aircraft engine remaining useful life estimation via a double attention-based data-driven architecture. *Reliab. Eng. Syst. Saf.* **2022**, *221*, 108330. [CrossRef]
25. Ke, G.; He, D.; Liu, T.Y. Rethinking positional encoding in language pre-training. *arXiv* **2020**, arXiv:2006.15595
26. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.
27. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv* **2022**, arXiv:2210.02186.
28. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
29. Li, Y.; Nilkitsaranont, P. Gas turbine performance prognostic for condition-based maintenance. *Appl. Energy* **2009**, *86*, 2152–2161. [CrossRef]
30. Myers, L.; Sirois, M.J. Spearman correlation coefficients, differences between. *Encycl. Stat. Sci.* **2004**, *12*. [CrossRef]
31. Zhang, C.; Lim, P.; Qin, A.K.; Tan, K.C. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2306–2318. [CrossRef] [PubMed]
32. Moore, D.S. *Introduction to the Practice of Statistics*; WH Freeman and Company: New York, NY, USA, 2009.
33. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
34. Li, W.; Zhang, Z.; Wang, X.; Luo, P. Adax: Adaptive gradient descent with exponential long term memory. *arXiv* **2020**, arXiv:2004.09740.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.