Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 75–80
DOI: 10.15439/2025F8176
ISSN 2300-5963 ACSIS, Vol. 44

Interpreting NAS-Optimized Transformer Models for Remaining Useful Life Prediction Using Gradient Explainer

Messaouda Nekkaa
ORCID: 0000-0002-6472-8266
University M'hamed Bougara of Boumerdes
LIST / Electrical Systems Engineering Department
35000 Boumerdes, Algeria
Email: m.nekkaa@univ-boumerdes.dz

Mohamed Abdouni
Sonatrach Industry
Djenane El Malik, Hydra, 16111 Algiers, Algeria
Email:mohamed.abdouni@sonatrach.dz

Dalila Boughaci
ORCID: 0000-0001-5210-8951
University of Science and Technology Houari Boumediene
LRIA / Computer Science Department
BP 32 El-Alia, Bab Ezzouar, 16111 Algiers, Algeria
Email: dalila_info@yahoo.fr, dboughaci@usthb.dz

***Abstract*—Remaining Useful Life (RUL) estimation of complex machinery is critical for optimizing maintenance schedules and preventing unexpected failures in safety-critical systems. While Transformer architecture has recently achieved state-of-the-art performance on RUL benchmarks, their design often relies on expert tuning or costly Neural Architecture Search (NAS), and their predictions remain opaque to end users. In this work, we integrate a Transformer whose hyperparameters were discovered via evolutionary NAS with a gradient-based explainability method to deliver both high accuracy and transparent, per-prediction insights. Specifically, we adapt the Gradient Explainer algorithm to produce global and local importance scores for each sensor in the C-MAPSS FD001 turbofan dataset. Our analysis shows that the sensors identified as most influential, such as key temperature and pressure measurements, match domain-expert expectations. By illuminating the internal decision process of a complex, NAS-derived model, this study paves the way for trustworthy adoption of advanced deep-learning prognostics in industrial settings.**

***Index Terms*—Remaining Useful Life (RUL), Transformers, Neural Architecture Search (NAS), Explainable AI (XAI), Gradient Explainer, C-MAPSS, Interpretability.**

I. Introduction

PROGNOSTICS and Health Management (PHM) plays a critical role in modern industrial systems, enabling increased reliability, optimized maintenance, and the prevention of catastrophic failures in high-value assets such as aircraft engines and manufacturing equipment [1]. A core component of PHM is the accurate estimation of Remaining Useful Life (RUL), the time before a component or system can no longer perform its intended function.

The rise of deep learning has significantly advanced RUL prediction. Recurrent Neural Networks (RNNs) [2], and more recently Transformer-based architectures [3], have demonstrated strong performance due to their ability to model complex temporal dependencies in multivariate sensor data.

Building on these advances, Mo Hyunho et al. [4] proposed a Neural Architecture Search (NAS) framework using evolutionary algorithms to automatically discover optimal Transformer architectures for RUL prediction. Applied to the well-established C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset [6], their NAS-derived Transformers outperformed manually designed alternatives, setting a new performance benchmark [4].

Despite these gains, deep-learning complex models often operate as "black boxes" [7]. Their complex, high-dimensional structures obscure the reasoning behind predictions.

In safety-critical settings, this lack of interpretability is a major barrier to adoption, where understanding why a model predicted a specific RUL is essential for trust, verification, and regulatory acceptance.

Explainable AI (XAI) seeks to address this issue by providing human-understandable insights into model behavior. However, most existing XAI studies focus on standard or simpler architectures, leaving the interpretability of NAS-derived Transformers underexplored, especially within the PHM domain [7], [8].

To our knowledge, no prior work has applied advanced gradient-based XAI techniques to these automatically discovered architectures in the context of RUL estimation.

This paper addresses that gap by adapting SHAP's Gradient Explainer ; a theoretically grounded and computationally efficient method ; for use with the NAS-optimized Transformer developed by Mo Hyunho et al. Our goal is to enhance the

Topical area: Advanced Artificial Intelligence in Applications

transparency of this state-of-the-art model by generating global and local feature attributions for RUL predictions on the C-MAPSS FD001 subset.

Our contributions are threefold:

1) **Gradient-based Explanation for NAS-Transformer:** We adapt and apply the Gradient Explainer algorithm to a NAS-optimized Transformer architecture specifically designed for RUL prediction.
2) **Global and Local Attribution Analysis:** We perform comprehensive explanation analysis, including both global sensor rankings and per-instance local saliency maps, on the FD001 subset of C-MAPSS.
3) **Actionable Insights for PHM:** We extract interpretable, domain-relevant insights into which sensors and time points most influence the model's predictions, enhancing trust, transparency, and deployability in industrial contexts.

The rest of this paper is organized as follows:

Section II reviews related work on RUL prediction and explainable AI. Section III describes the dataset, model architecture, and the adaptation of the Gradient Explainer. Section IV presents experimental results, including global and local explanations. Section V concludes with future research directions.

II. Related Work

This section reviews literature pertinent to our research, covering Remaining Useful Life (RUL) prediction with deep learning, the role of Neural Architecture Search (NAS) in Prognostics and Health Management (PHM), existing Explainable AI (XAI) techniques for complex models, and the specific challenges and advancements in explaining Transformer and NAS-optimized architectures.

A. RUL Prediction in PHM

Remaining Useful Life (RUL) refers to the time remaining before a system fails, expressed as RUL $= T - t$, where T is the failure time and t is the current time [1]. RUL estimation methods are broadly categorized into model-based and data-driven approaches. Model-based methods rely on prior physical knowledge, which can be hard to generalize in practice and may struggle with the complexities of real-world degradation processes. In contrast, data-driven approaches, particularly those leveraging deep learning (DL), have gained prominence due to their ability to learn complex patterns directly from sensor data and enabling end-to-end modeling, eliminating the need for manual feature engineering [2].

Early DL applications in RUL prediction included Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), which showed promise in feature extraction from time-series data. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, became popular for their inherent ability to model temporal dependencies in sequential sensor readings. However, RNNs can face challenges with long range dependencies and computational efficiency for long sequences [9], [10].

B. Transformer-Based Models for Time Series

Transformer architecture, originally introduced for natural language processing in the famous paper of Vaswani et al. [3], has emerged as a powerful self-attention mechanism that allows it to capture global dependencies between input sequence elements effectively, overcoming some limitations of RNNs. Consequently, Transformers have been increasingly adapted for various time-series forecasting tasks, including RUL prediction, often demonstrating superior performance.

C. Neural Architecture Search (NAS) in Deep Learning

While DL models, including Transformers, offer significant potential, their performance is highly dependent on their architecture. Designing optimal architecture manually is a time-consuming, iterative, and expertise-driven process [4]. Neural Architecture Search (NAS) has emerged as a field that automates this design process, algorithmically searching for the best-performing neural network architecture for a given task and dataset [4].

D. Explainable AI (XAI) for Complex Models

The increasing complexity and performance of DL models, especially transformer-based models with their attention characteristics, often come at the cost of interpretability, leading to their characterization as "black boxes". In safety-critical applications like PHM, this lack of transparency is a major concern, as understanding why a model makes a certain prediction is crucial for trust, debugging, and regulatory compliance. Explainable AI (XAI) encompasses a range of techniques aimed at making the decisions of AI systems more understandable to humans [11].

Common XAI methods can be broadly categorized. Perturbation-based methods, like LIME (Local Interpretable Model-agnostic Explanations), explain individual predictions by learning a simpler, interpretable model on local perturbations of the input [11].

Surrogate models aim to approximate the complex model with a more transparent one. Gradient-based methods, such as Integrated Gradients and SmoothGrad, utilize model gradients to attribute importance to input features. SHAP (SHapley Additive exPlanations), grounded in co-operative game theory, provides a unified framework for feature attribution by calculating Shapley values, which represent the marginal contribution of each feature to the prediction [12].

III. Material and methods

In this section, we present our methodological frame-work. We first describe the C-MAPSS FD001 dataset and its preprocessing pipeline. Next, we introduce the NAS-optimized Transformer architecture used for RUL predic-tion. Finally, we detail our adaptation of the SHAP Gradient Explainer for feature-attribution analysis applied to this model.

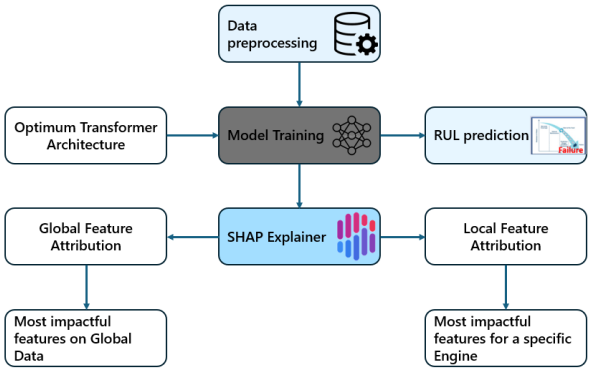

Fig. 1. Methodological Frame-Work

A. Data and Preprocessing

We base our experiments on NASA's widely used C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset, which simulates turbofan engine degradation under different operating conditions and fault modes. C-MAPSS comprises four subsets (FD001–FD004), each containing multivariate time-series from 21 sensors and 3 operating settings.

In this work, we focus on FD001, which models a single fault mode under one operating condition [5].

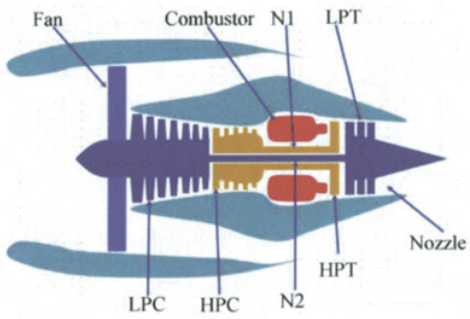

Fig. 2. Diagram of the turbofan Engine

Data preprocessing steps were aligned with those typically employed for this dataset and consistent with the foundational work [4]:

- **Sensor Selection:** From the original 21 sensor channels, we computed the 21×21 inter-sensor Pearson correlation matrix to identify constant or redundant signals. Any sensor with zero variance (constant readings) or entirely null values was removed, leaving 14 informative sensors.
- **Normalization:** All sensor and aggregate features were scaled to [0, 1] using min–max normalization, with scaling parameters fitted exclusively on the FD001 training set to avoid data leakage.
- **Windowing:** We applied a sliding window of 40 raw timesteps and appended 2 aggregate rows (slope and mean), resulting in 42-timestep sequences. The target RUL is defined as the number of cycles remaining at the final point in each window.

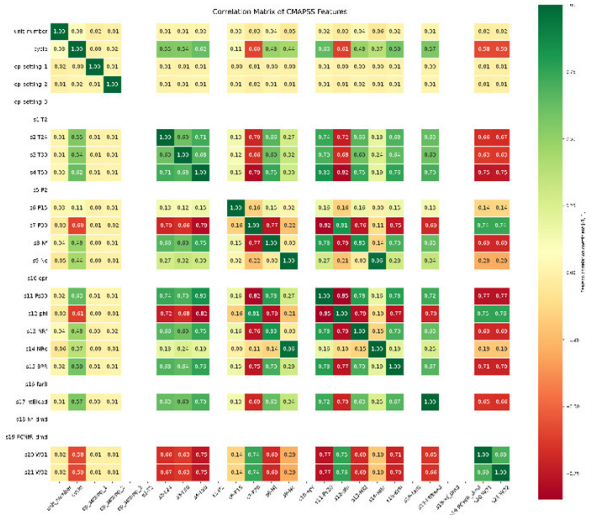

Fig. 3. Pearson's correlation matrix heat map of the Commercial Modular Aero-Propulsion System

B. Foundational NAS-Optimized Transformer Architecture

Our work builds upon the Transformer architecture developed by Mo Hyunho et al. [4], who applied Neural Architecture Search (NAS) to design high-performing models for RUL prediction. Rather than re-running their computationally intensive search process, we adopt the optimal architecture they identified as the basis for our explainability study.

This architecture was discovered using an evolutionary algorithm that explored an 11-dimensional genotype defining various hyperparameters of the Transformer, including embedding dimensions, number of attention heads, feed-forward layer dimensions, and the number of encoder/decoder layers.

The core structure of this NAS-optimized Transformer architecture, as described by Mo Hyunho et al. [4], features several key components tailored for time-series RUL prediction:

- **Input Representation:** Each input is a multivariate time-series window with 42 timesteps and 14 sensor channels, resulting in an input matrix of shape (42, 14). The 42 timesteps include 40 raw cycles and 2 aggregate features (slope and mean), as described in Section III-A.
- **Embedding and Positional Encoding:** Raw sensor readings at each timestep are first passed through an input embedding layer to project them into a higher-dimensional space (`d_model`) . To retain temporal information, sinusoidal positional encodings are added to these embeddings.
- **Dual-Encoder Mechanism:** A key feature of the architecture is its use of two parallel encoders:
  - *A Sensor Encoder:* that applies multi-head self-attention across the sensor dimension to assess inter-sensor dependencies.

– *A Timestep Encoder:* that uses self-attention across the time dimension to capture temporal patterns.
Each encoder is composed of `N_enc` layers, each containing multi-head attention and position-wise feed-forward sublayers, combined with residual connections and layer normalization.

- **Feature Fusion:**
Outputs from the sensor and timestep encoders—denoted F_s and F_t are concatenated and passed through a fusion layer:

$$\text{Fusion}(F_s, F_t) = \text{Concat}(F_s, F_t) \cdot W^F \quad (1)$$

This operation merges sensor-wise and temporal features into a unified representation.
- **Decoder:** The fused features are input to a decoder composed of N_{dec} layers, again using multi-head attention and feed-forward sublayers. The decoder processes only the final α timesteps of the encoder output $typically\alpha = 4$, focusing on recent history for prediction. Its final output is a scalar representing the estimated RUL.

We configured our model using the specific optimal genotype parameters reported by Mo Hyunho et al. [13], ensuring consistency with the NAS-discovered Transformer architecture used in their original work.

C. Gradient Explainer Algorithm

To interpret the predictions of the NAS-optimized Transformer, we adopted SHAP's Gradient Explainer [12], a member of the gradient-based attribution family introduced in Section II. Gradient Explainer estimates feature contributions by computing expected gradients relative to a background distribution, enabling both local explanations (per Engine) and global insights (across the dataset).

This method was chosen for its compatibility with non-standard architecture Transformers and structured multivariate time series, as encountered in our 42×14 input windows. While other techniques such as LIME and Integrated Gradients are valuable in broader explainability contexts [14], [15], SHAP Gradient Explainer offers theoretical consistency, computational efficiency, and additive attribution, aligning well with the goals of transparency in RUL forecasting.

1) Background Selection: : SHAP requires a background dataset to serve as a reference for calculating expected gradients. We use all 100 training windows as the background set, ensuring full coverage of operating conditions and RUL states. This choice balances computational efficiency with stability in the resulting attributions.

2) Batched SHAP Computation.: Due to memory constraints, SHAP values are computed in batches of size 10. Each test sample (of shape 42×14) is passed to the explainer, which returns a tensor of SHAP values with the same shape. These represent the contribution of each sensor at each timestep (including slope and mean rows) to the model's RUL prediction.

IV. Results and Discussion

This section presents the results of our explainability pipeline to evaluate the NAS-optimized. We report results on global feature importance, local attribution for specific predictions, and validate the reliability of the explanations through coherence checks.

1) Global Attribution.: To understand which features were most influential across all test samples, we applied SHAP's Gradient Explainer using 100 stratified background windows. The resulting SHAP values were aggregated across all test inputs, and the top features were visualized using a bar summary plot (Figure 4) and the beeswarm summary plot (Figure 5).

The most impactful feature was `BPR_t41`, the mean value of the Bypass Ratio sensor, which positively influenced RUL predictions. Other highly influential features included the slopes of `phi`, `P30`, and the mean or trend of sensors like `T24` and `W32`. These results confirm that both recent degradation trends (slope features) and operating-level signals (mean features) contribute meaningfully to the model's decisions.

The top-ranked sensors correspond to known degradation-related physical components, supporting the model's alignment with domain expectations. Less informative features were grouped into an "Other" category, highlighting the concentration of decision impact among a small subset of sensor-time features.

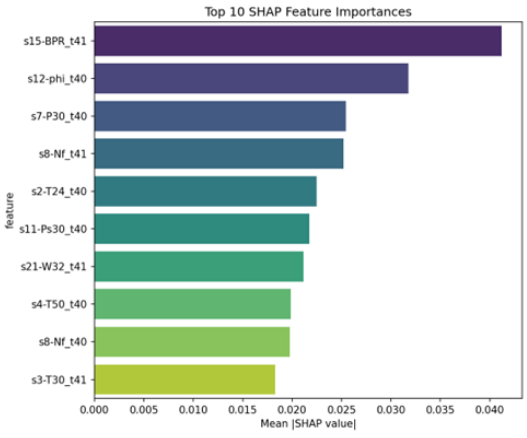

Fig. 4. Global Sensor Ranking barplot

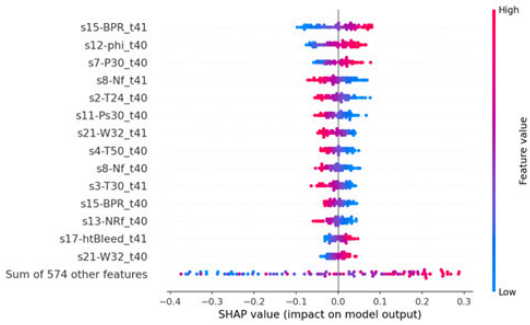

Fig. 5. Global SHAP Summary (Beeswarm) Plot

2) Local Feature Attribution: : To explore how the model forms individual predictions, we examined SHAP waterfall plots for representative test samples. Figure 6 shows a case where the predicted Remaining Useful Life (RUL) was significantly lower than average (0.072 vs. 0.709). Negative contributions came from slope features such as `phi_t40`, `NRf_t40`, and `P30_t40`, which indicate rapid degradation in pressure and rotational speed. A single feature, `BPR_t41`, contributed positively, but only marginally.

Notably, the largest reduction in prediction came from the aggregate contribution of 579 other features, which collectively pulled the estimate downward by -0.25. This highlights the model's ability to synthesize both prominent and subtle signals across the input sequence. The explanation aligns with real-world intuition: sharp declines in critical sensors indicate worsening engine health, justifying a lower RUL forecast.

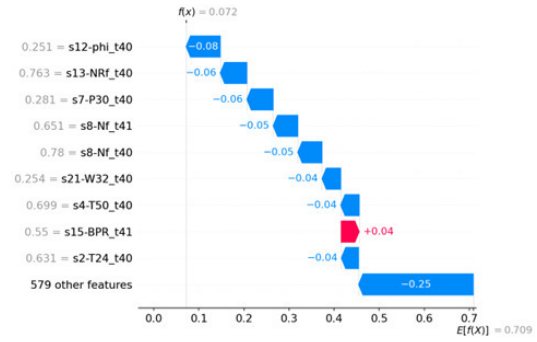

Fig. 6. Local SHAP Waterfall Plot Engine 41

A. Coherence Checks

To assess the trustworthiness of the model's explanations, we conducted a qualitative analysis of the SHAP outputs. Specifically, we reviewed whether the top-ranked features identified by the Gradient Explainer aligned with known degradation indicators in the turbofan engine domain.

Our global attribution analysis revealed that the most influential features included trends and mean values from key sensors such as Bypass Ratio (BPR), high-pressure compressor pressure (P30), rotational speeds (`Nf`, `NRf`), and temperatures (`T24`, `T30`). These are consistent with established knowledge about engine wear and failure modes. Similarly, local explanations for individual predictions showed that decreasing trends in these features often led to lower RUL estimates, reinforcing their interpretability.

Although we did not formally quantify explanation robustness (e.g., using Spearman correlation), the consistent emergence of domain-relevant features in both global and local attributions suggests that the model has learned meaningful and physically plausible relationships. This coherence is a promising indicator for the model's transparency and practical applicability in industrial settings.

V. Conclusion

This paper presents an explainability study of a NAS-optimized Transformer model for Remaining Useful Life (RUL) prediction on the C-MAPSS FD001 benchmark. We integrate SHAP's Gradient Explainer into the model pipeline to generate both global sensor importance rankings and local per-sample attribution maps. Our results show that the model's most influential features, particularly sensor trends and means in airflow, pressure, and temperature, are consistent with known degradation indicators in jet engines.

By illuminating how the model forms each prediction, our approach enhances transparency and supports trust in deep learning-based prognostics. While this study focuses on a single dataset and architecture, the method is generalizable and can be extended to other PHM tasks or architectures.

Future work will incorporate formal stability tests, expert validation, and broader dataset coverage.

References

[1] E. Zio, "Prognostics and health management (PHM): Where are we and where do we (need to) go in theory and practice," *Rel. Eng. Syst. Saf.*, vol. 218, Art. 108119, 2022. Available: https://doi.org/10.1016/j.ress.2021.108119.

[2] O. Serradilla, E. Zugasti, J. Rodriguez, *et al.*, "Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects," *Appl. Intell.*, vol. 52, pp. 10934–10964, 2022. Available: https://doi.org/10.1007/s10489-021-03004-y

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. Available: https://doi.org/10.48550/arXiv.1706.03762

[4] H. Mo and G. Iacca, "Evolutionary neural architecture search on transformers for remaining useful life prediction," *Mater. Manuf. Process.*, pp. 1–18, 2023. Available: https://doi.org/10.1080/10426914.2023.2199499

[5] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," *Proc. Int. Conf. Prognostics Health Manag. (PHM)*, Denver, CO, USA, pp. 1–9, 2008. Available: https://doi.org/10.1109/PHM.2008.4711414.

[6] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, pp. 45–74, 2024. Available: https://doi.org/10.1007/s12559-023-10179-8

[7] A. T. Keleko, B. Kamsu-Foguem, R. H. Ngouna, and A. Tongne, "Health condition monitoring of a complex hydraulic system using deep neural network and DeepSHAP explainable XAI," *Adv. Eng. Softw.*, vol. 175, Art. 103339, Jan. 2023. Available: https://doi.org/10.1016/j.advengsoft.2022.103339

[8] G. Youness and A. Aalah, "An explainable artificial intelligence approach for remaining useful life prediction," *Aerospace*, vol. 10, no. 5, pp. 1–23, 2023. Available: https://doi.org/10.3390/aerospace10050474

[9] T. Markovic, A. Dehlaghi-Ghadim, M. Leon, A. Balador, and S. Punnekkat, "Time-series anomaly detection and classification with long short-term memory network on industrial manufacturing systems," *Proc. 18th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 35, pp. 171–181, 2023. Available: https://doi.org/10.15439/2023F5263.

[10] S. Zhao, Y. Zhang, S. Wang, B. Zhou, and C. Cheng, "A recurrent neural network approach for remaining useful life prediction utilizing a novel trend features construction method," *Measurement*, vol. 146, pp. 279–288, 2019. Available: https://doi.org/10.1016/j.measurement.2019.06.004.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[12] SHAP (SHapley Additive exPlanations): a game theoretic approach to explain the output of any machine learning model. Available: https://github.com/shap/shap

[13] M. Ho, "NAS_transformer: Neural architecture search for transformer-based models," GitHub repository, 2023. Available: https://github.com/mohyunho/NAS_transformer

[14] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," *Proc. IEEE SmartWorld*, San Francisco, CA, USA, pp. 1–6, 2017. Available: https://doi.org/10.1109/UIC-ATC.2017.8397411.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2017. Available: https://doi.org/10.48550/arXiv.1703.01365.