National College of Ireland

Feature-Focused Defense: The Role of Attention Mechanisms in Enhancing Adversarial Attack Detection

MSc Research Project
Master of Science in Data Analytics

Mahesh Ryada
Student ID: 23318686

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Mahesh Ryada

Student ID: 23318686

Programme: Master of Science in Data Analytics **Year:** 2026

Module: MSc Research Practicum Part 2

Supervisor: Abdul Qayum

Submission Due Date: 28/01/2026

Project Title: Feature-Focused Defense: The Role of Attention Mechanisms in Enhancing Adversarial Attack Detection

Word Count: 7816 **Page Count:** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mahesh Ryada

Date: 28/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Feature-Focused Defence: The Role of Attention Mechanisms in Enhancing Adversarial Attack Detection

Mahesh Ryada

X23318686

Abstract

Adversarial manipulation of sensor data poses a major challenge in safety-critical industrial environments, where even subtle perturbations can disrupt predictive maintenance models and lead to unreliable engine health assessment. Traditional machine learning and deep learning approaches used in such applications remain vulnerable to these perturbations, as their fixed feature-weighting strategies and static representations struggle to capture the nuanced deviations introduced by adversarial interference. Existing defense` strategies—including adversarial training, feature-space regularization, and domain adaptation—show performance gains but often fail to generalize across varying attack scenarios, leading to reduced robustness in dynamic sensor conditions. This study investigates a feature-tokenized transformer architecture that leverages attention-based feature interaction modelling to strengthen adversarial detection in turbofan engine sensor data. The model dynamically recalculates feature relevance, enabling improved identification of subtle inconsistencies in manipulated inputs when compared with conventional architectures. A comprehensive evaluation across multiple baseline machine learning and deep learning models demonstrates that the transformer-based approach achieves the strongest performance, recording an accuracy of **0.93** with substantial improvements in precision (0.95), and F1-score (0.87). The feature-level attention mechanism contributes to stable decision boundaries even under perturbed conditions, enabling significantly better discrimination between clean and adversarial samples. These findings highlight the suitability of attention-driven architectures for adversarial resilience in sensor-driven predictive maintenance and point toward their potential for enhancing the reliability of industrial monitoring systems exposed to manipulation risks.

Keywords: Adversarial Attack Detection, Attention Mechanisms, FT-Transformer, Predictive Maintenance.

1 Introduction

1.1 Problem Background

Sensory stream adversarial manipulation is emerging as a significant reliability threat as predictive models gain increasingly important roles in industrial decision making. Even minor, well-designed modifications to sensor measurements can cause distortion of prognostic predictions and cause wrong maintenance interventions in safety-critical environments like aircraft engine health monitoring, and misclassify faults or incorrectly predict the remaining useful life. Previous research demonstrates machine learning and deep

learning models to be significantly vulnerable to adversarial attacks, samples that are adversarial perturbed significantly diminish detection and prediction accuracy (Anasuri, 2022; Zhao et al., 2025). With the continuous transition to condition monitoring based on data-driven pipelines, the robustness to adversarial interference is becoming a crucial factor in the operational safety and the system reliability.

The motivation of this work is the repeated occurrence of adversarial vulnerability in cyber-physical and industrial analytics. Soft-sensor models in manufacturing, chemical plants, and energy systems tend to be unstable to adversarial noise due to the commonly used learning goals being accuracy at nominal conditions and not resilience to distribution shift (Xie et al., 2023; Guo et al., 2024). Similar weaknesses can be found in the case of intrusion detection and IoT monitoring, where both neural and tree-based models fail in the situation when attackers alter feature distribution to avoid detection (Vitorino et al., 2023; Gaber et al., 2025). These trends indicate that more detailed inter-feature structure is required and that the ability to detect subtle, coordinated deviations that are indicative of adversarial behaviour is needed.

Various adversarial training, perturbation-informed feature learning, domain adaptation, and data augmentation through GANs have been explored. As an example, attention-targeting attacks have been addressed by proposing attention-guided strategies (Zhao et al., 2025), and latent-space adversarial training has enhanced robustness of industrial soft sensors (Xie et al., 2023). Adversarial augmentation has demonstrated benefits in cybersecurity settings, usually at the expense of overfitting and poor generalization to previously unseen attack patterns (Vitorino et al., 2023; Gaber et al., 2025). It has also been reported that efforts to stabilize adversarial training with historical gradient information have been made, but the accurate modelling of adversarial signatures in nonstationary settings is still difficult to achieve (Guo et al., 2024). In general, it is found in the literature that there are numerous traditional and deep learning solutions, which are not as flexible as is needed to ensure reliability in cases where adversarial changes alter sensor behaviour.

1.2 Research Question

The core research question guiding this study is:
"How effectively feature-level attention mechanisms can be integrated with deep learning architectures to detect adversarial attacks better than traditional machine learning algorithms in terms of predictive performance and robustness?"

Justification
Neural networks with traditional machine learning and deep learning algorithms tend to become unstable when they are subject to adversarial perturbations, and this is in part due to the fact that their fixed feature-weighting components are not able to represent the fine-scale changes that are added during manipulation. Adversarial training and domain adaptation can enhance resilience, but often they are not sufficiently flexible to be applicable to diverse and unobservable attack patterns. This weakness has led to the need of architectures that can

analyse feature interactions dynamically instead of depending on static representations. This is made possible by feature-level attention mechanisms, which re-computes the significance of each feature on each input, giving the model the ability to identify irregular or abnormal relationships in sensor data that adversarial attacks tend to exploit. The effectiveness of such attention mechanisms to be incorporated into deep learning models and whether it can be more effective and robust in predictive performance than other traditional machine learning methods is an important question to answer to enhance adversarial detection in sensor-based industrial systems (Xie et al., 2023; Guo et al., 2024).

1.3 Contribution of the Study

The current study is valuable because it assesses a feature-tokenized transformer based adversarial attack detector on turbofan engine sensor data and how it outperforms the traditional machine learning and deep learning algorithms. The attention mechanism is built in the FT-Transformer that allows dynamic evaluation of the feature interactions and facilitates more credible detection of adversarial deviations in engine measurements. The research confirms that this architecture can be largely used to enhance accuracy, recall, and robustness in general, and it provides valuable information on how sensor-driven predictive maintenance and industrial monitoring systems can be improved in terms of security and robustness.

1.4 Report Organization

The present research is useful as it evaluates a feature-tokenized transformer-based adversarial attack detector on data presented by turbofan engine sensors and its superiority over the conventional machine learning and the deep learning algorithms. The FT-Transformer constructs the attention mechanism enabling dynamic analysis of the feature interactions and enables more plausible detection of the adversarial deviations on the measurements of the engine. It is established that this architecture can be used to a great extent to improve accuracy, recall, and robustness in general; the research can also give useful information on how sensor-driven predictive maintenance and industrial monitoring systems can be made more secure and robust.

2 Literature Review

The literature review has noted that both the traditional machine learning and deep learning models have incurable vulnerabilities when subjected to adversarial perturbation, especially with fixed feature weighting and a lack of flexibility.

2.1 Machine Learning Models for Security-Critical Applications

The fact that machine learning (ML) models are capable of learning patterns in complex, high-dimensional data has led to their use in cybersecurity, industrial monitoring, healthcare

analytics, and network intrusion detection. Conventional ML models like the Random Forest (RF), Support Vector Classifier (SVC), Logistic Regression (LR), and Decision Trees (DT) are highly predictive in the structured or tabular data framework, especially in network traffic analytics and industrial sensor systems. These models are strong in interpretability and simplicity to implement, but have problems in the environments with noisy, incomplete, and highly imbalanced data. Research indicates that supervised ML models can be easily attacked by adversarial perturbed samples, which result in extreme deterioration of performance in intrusion detection and malware classification tasks (Alshahrani et al., 2022; Alzaidy and Binsalleeh, 2024).

A variety of works has tried to enhance the robustness of ML by adversarial training, noise-resilient feature extraction, and ensemble learning. As an example, methods that use GAN-generated adversarial traffic achieve much better resilience of ML-based DdoS detection models (Mustapha et al., 2023; Gaber et al., 2025). Likewise, the optimization of ML algorithms like CatBoost to identify intrusion demonstrates it has better performance in terms of classifying harder attacks, but the vulnerability remains in the advanced adversarial environment. Such developments notwithstanding, one of the most common issues is that the baseline ML models are not inherently robust to gradient-based adversarial attacks (e.g., FGSM, PGD, JSMA), which reveal inherent vulnerabilities in their generalization and decision-boundary stability (Haroon and Ali, 2022).

Moreover, machine learning models are not effective in practice when used in physical systems, which include IoT, WSNs, industrial sensors, and power grids, because noise, anomalies, or missing values can cause a poor quality of data to be used (Shihab et al., 2024; Albattah and Rassam, 2023). As Owezarski (2023) illustrates, the use of Random Forest, which is commonly regarded as a powerful baseline, can be easily evaded with small and targeted perturbation. Hence, the literature review indicates that more robust feature representation and training methods are urgently required to be resistant to adversarial manipulation in a variety of application areas.

2.2 Deep Learning Models and Their Role in Enhancing Robustness

Traditional ML has been outperformed by deep learning (DL) models, such as CNNs, RNNs, LSTMs, autoencoders and deep transformer-based architectures, in complex and unstructured settings. They can identify hierarchical feature representations, which is why they are used in anomaly detection of industrial systems, malware analysis, IoT cybersecurity, and power grid monitoring (Sauka et al., 2022; Li et al., 2023). Nevertheless, they are complex, which makes them vulnerable: deep networks are vulnerable to adversarial examples leading to misclassification with imperceptible perturbations. A number of studies have also investigated the weaknesses of DNNs during adversarial scenarios, particularly in areas of critical decision-making (i.e., power system monitoring and healthcare) (Huang and Li, 2022; Albattah and Rassam, 2023).

More recent advances made to make DL robust are adversarial training models, distillation-trained defences, causal-based mitigation, and feature-space adversarial learning. As an illustration, Zhao et al. (2025) presented attention-guided defense modules, specifically, FPAS and ANL, to combat attention-shifting and attention-attenuation attacks, respectively, and attained substantial enhancement of adversarial robustness with minimum intrusion of the model. Similarly, Xie et al. (2023) applied adversarial perturbations to the latent space of a deep supervised variational autoencoder (DSVAE) to enhance the soft sensor reliability of noisy industrial settings. To stabilize gradient estimation and minimize robust overfitting, domain-adaptive adversarial training has also been investigated to stabilize gradient estimation and in the case of deep learning-based soft sensors (Guo et al., 2024), it has proven effective.

DL-based DdoS detectors (Mustapha et al., 2023), healthcare anomaly detectors based on ConvLSTM (Albattah and Rassam, 2023), and GAN-enhanced intrusion detectors (Gaber et al., 2025) have demonstrated different levels of effectiveness in cybersecurity applications. However, even these systems have problems with remaining resilient to white-box as well as black-box adversarial examples like FGSM, PGD, DeepFool, and C&W (Anasuri, 2022; Alzaidy and Binsalleeh, 2024). Despite being more accurate than classical ML algorithms, the high vulnerability to adversarial manipulation has been an Issue with the DL models.

2.3 Advanced and Hybrid Models: Strengths, Limitations, and Efficiency Concerns

Advanced hybrid models—such as GAN-integrated IDS, domain-adaptive adversarial training frameworks, attention-based architectures, and optimized ensembles—have been proposed to outperform standard ML/DL approaches. Techniques like WCSAN-PSO (Barik et al., 2024), GAN-based dataset augmentation (Gaber et al., 2025), and historical gradient-based adversarial attack generation (Guo et al., 2024) achieve exceptional detection performance, often surpassing baseline models in adversarial environments. Transformer-based approaches and feature pyramid architectures, as introduced by Zhao et al. (2025), also show strong defense capabilities against subtle attention-based adversarial shifts.

However, these methods often require **heavy memory, substantial computational resources, and long training times**, making them unsuitable for real-time IoT, WSN, or industrial monitoring systems that operate under strict resource constraints. Many hybrid and adversarial trained systems involve iterative gradient generation, domain adaptation, or GAN-based sample synthesis, which further increases model complexity and deployment cost. Studies such as those by Vitorino et al. (2023) and Shihab et al. (2024) note that high-performing models may lack scalability, making them impractical for edge or embedded environments. Additionally, methods like adversarial distillation, multi-stage training, or deep autoencoder stacking (Xie et al., 2023; Alzaidy & Binsalleeh, 2024) may improve robustness but at the expense of latency and energy efficiency.

Table 1. Summary of Related Studies and Their Relevance to FT-Transformer

Author(s)	Work Summary	Model Used	Contribution	Limitations
Zhao et al. (2025)	Developed universal attention-guided defense mechanisms.	Attention-based FPAS, ANL	Improved robustness against attention-shift attacks.	High computational cost; not optimized for tabular sensor data.
Vitorino et al. (2023)	Explored adversarial realism for IoT intrusion detection.	DL-based IDS	Enhanced classification accuracy on IoT attack data.	Overfitting to specific perturbation types; struggles with tabular sensor heterogeneity.
Owezarski (2023)	Investigated adversarial vulnerability of Random Forest IDS.	Random Forest	Highlighted brittleness of tree models under small perturbations.	RF fails to detect subtle adversarial deviations in feature interactions.
Gaber et al. (2025)	Designed GAN-based adversarial defense for flying-IoT systems.	GAN + Neural Networks	Improved detection via adversarial augmentation.	GAN training is unstable and computationally heavy.
Haroon & Ali (2022)	Studied adversarial training for IDSs.	ML-based IDS	Increased robustness against gradient-based attacks.	High dependency on attack type; lacks dynamic feature weighting.
Sauka et al. (2022)	Developed robust and explainable IDS using DL.	Deep neural networks	Improved detection and interpretability.	Vulnerable to subtle adversarial feature manipulations.
Albattah & Rassam (2023)	Examined adversarial attacks on hybrid ConvLSTM for	Hybrid ConvLSTM	Better temporal anomaly detection.	Inefficient for static tabular data; sensitive to input perturbations.

	healthcare.			
Anasuri (2022)	Surveyed adversarial attacks and defences in deep neural networks.	DNN architectures	Provided foundational insights into vulnerabilities.	Highlights universal DL weakness: no inherent feature-interaction adaptability.
Mustapha et al. (2023)	Proposed adversarial ANN for DdoS detection.	Adversarial Neural Network	Strengthened DdoS classification accuracy.	Not designed for sensor-driven industrial datasets with fine-grained features.
Alzaidy & Binsalleeh (2024)	Assessed adversarial effects in malware DL systems.	CNN & RNN models	Enhanced robustness with hybrid defences.	Sequence-based architectures poorly suited for tabular sensor data.
Huang & Li (2022)	Introduced mitigation strategy for ML-based power system IDS.	ML-based classifiers	Reduced attack influence on power grid detection.	Relies on fixed feature weights; lacks adaptive attention.

Table 1 provides a brief comparison of the related studies which are the most important, and demonstrates how each of them informs or supports the application of FT-Transformer to the tasks with tabular data.

2.4 Research Gaps and Motivation for FT-Transformer

2.4.1 Drawbacks in Existing Works

Although there is a lot of research, a number of constraints continue to exist in the traditional ML, deep learning, and hybrid adversarial defense methods. To begin with, numerous adversarial training methods have strong overfitting, gradient instability, or fail to generalize to previously unknown types of attacks (Guo et al., 2024; Haroon and Ali, 2022). Second, the current defense systems are based on computationally intensive architectures or multi-stage training pipelines that cannot be deployed in real-time (Barik et al., 2024; Zhao et al., 2025). Third, ML and DL models often do not inherently support tabular, mixed-type, or incomplete data that is present in industrial and IoT data, and thus they become brittle to noise, missing values, or a change in distribution (Xie et al., 2023; Shihab et al., 2024). Moreover, models that use GAN-based adversarial generation or gradient-based perturbation models usually need to be optimized with complicated methods and lack interpretability, which further makes them difficult to implement (Gaber et al., 2025; Alzaidy and Binsalleeh, 2024). Throughout all the fields, lack of interaction between features, unstable training under

adversarial environment, and absence of universal defence mechanisms are all common problems in the literature.

2.4.2 How FT-Transformer Can Address These Drawbacks

FT-Transformer provides a robust gateway to fill these gaps since its architecture is specifically made to process tabular data with high efficiency, strong feature interaction modelling and transformer-based attention mechanisms that boosts robustness. In comparison to deep CNNs, LSTMs, or GAN-hybrid systems, FT-Transformer needs much less computational power and can use sparse attention, which allows it to conduct inference much faster and is applicable to the inspection tasks of IoT, WSN, and industrial systems in real time. It has an innate ability to learn non-linear feature dependencies, which makes it resistant to adversarial perturbations in that it can stabilize feature relationships, which overcomes the challenges pointed out in domain-adaptive soft sensors (Guo et al., 2024) and attention-shift attacks (Zhao et al., 2025). Moreover, FT-Transformer requires no heavy adversarial retraining cycles, which can cause robust overfitting and computational overloads that occur during deep adversarial training models (Xie et al., 2023; Sauka et al., 2022).

Transformer-based attention combined with effective tabular learning FT-Transformer can provide lightweight, scalable, interpretable, and robust performance on industrial sensing, intrusion detection, and power system security monitoring. Its ability to stabilize feature embeddings and avoid deep gradient-based optimization endows it with potential to fill major weaknesses of adversarial robustness in the current literature.

3 Research Methodology

In this chapter, the methodological background of the study is described by providing the dataset, preprocessing pipeline, exploratory analysis, baseline comparisons, and the FT-Transformer framework applied to the detection of adversarial attacks in engine sensor data. All these steps led to the development of a systematic workflow, which led to the preparation of clean data, equal representation of classes, and meaningful analysis of features, which were then used to develop a model.

3.1 Model Motivation and Methodological Framework

In order to overcome the limitations of the existing works, this study will take advantage of the FT-Transformer model to overcome major limitations that have been witnessed in machine learning and deep learning approaches to adversarial attacks detection. Previous studies pointed to issues including the instability of feature interactions, noise sensitivity, poor generalization to adversarial perturbations, and high computational complexity of complicated architectures. These problems impact the accuracy of the models used to sensor-based industrial tasks, where adversarial examples may introduce very subtle corruption of the working measurements and misclassification. The reason why the FT-Transformer is used

in the current study is that the feature-tokenization mechanism and attention-based processing of the model allow to better capture the relationship between the input variables and become more resistant to the distortions caused by perturbations. This architecture is able to handle stable representation learning and be efficient, and hence it can be used in detecting adversarial attacks in turbofan engine data.

The flow of methodology used in this study is the CRISP-DM process model. The methodology starts with data comprehension, which is aimed at determining the pattern, form, and conduct that occurs in turbo fan engine parameters. The second step is data preparation, which is a process that entails cleaning, transformation, balancing, and scaling of data to ensure a uniform input. The modelling phase involves development of the FT-Transformer and benchmark model evaluation to compare with each other. The assessment phase looks at the classification results, loss behaviour and pattern of confusion matrix to establish the effectiveness of the model. Lastly, the results of these analyses are used to formulate the perspective of deployment, that is, how such detection systems can be incorporated into real-time adversarial monitoring systems.

3.2 Dataset Description

The data set utilized in this research includes engine operational readings, as well as adversarial stimuli in relation to predictive aircraft maintenance. It contains variables like operating setting values, corrected core shaft speed, speed variation, fuel-air characteristics, turbine bleed flow and valve status, and airflow at station 31, and two binary labels, which are engine health condition and adversarial manipulation status. The sample rows represent the format of the data, including values of such features as Corrected_Core_Speed_RPM, Core_Speed_Variation, and Airflow_at_Station in addition to health labels and indicators of adversarial attack. Each column gives information about engine behaviour, degradation behaviour and the effect of perturbations caused by attacks. The data set is complete and free of any missing or duplicate result, and this allows good preprocessing, analysis, and model development. Adversarial_Attack_Label is the target variable in the study, which determines the type of sample, clean or adversarial dataset.

3.3 Data Pre-processing

3.3.1 Data Integrity Checks

The preprocessing step was initiated by looking into the structural quality of the dataset to make sure that all the variables were fit to be analysed and developed into a model. This involved inspecting the format and nature of each column ensuring that all the features were captured in the right numerical format to be processed further. The data was reviewed on the presence of missing values, erroneous entries, and duplicate samples and none of the problems and structural errors were identified. This enabled the dataset to be transferred into feature preparation without any imputation or deletion process.

3.3.2 Feature Handling and Variable Refinement

After first checks, the focus on preparing the features in a form to be modelled was made. The data set included two label columns, one of them was referring to the state of the engine health, and the other was to adversarial manipulation. They were both verified to be binary so that they could be compatible with classification algorithms. Some of the variables that appeared to be redundant or did not make a significant contribution to the modelling goal were eliminated to form a more focused set of features. This refinement contributed to keeping a clean input space and minimized the amount of unwanted noise prior to model training.

3.3.3 Data Balancing Using SMOTE

The data set showed that there was a skew favouring clean samples and adversarial examples and this may result in biased learning to the majority. In order to cope with this problem, the Synthetic Minority Over-sampling Technique was used (SMOTE) on the training part of the data. The minority class is used to produce synthetic samples using interpolation between existing data points, producing a more evenly represented version of both classes. Consequently, the classifier received a balanced sample of adversarial and non-adversarial samples, which promoted fairer behaviour in learning during training.

3.3.4 Feature Scaling

After the dataset was balanced, the feature set was standardized to match the numerical ranges and to stabilize the model optimization. This transformation makes all numerical features the same scale so that no longer is the learning process dominated by attributes of larger magnitude. The processed data following scaling had homogeneous distribution patterns across features, which is ready to be used in both neural and transformer-based models to the extent of reliability.

3.4 EDA

The Exploratory Data Analysis was used in order to get an initial insight of the behaviour of engine measurements under normal and manipulated conditions. The visualizations were used to observe the imbalance in classes, the difference in important sensor values, and the patterns that were characteristic of clean and adversarial samples. The connections among features were investigated using distribution plots, scatter representation, and correlation visualization to determine trends and the possible impacts of the adversarial activity on engine parameters. Such observations helped in the later modelling by informing the subsequent stages of modelling on which variables were more sensitive to manipulation.

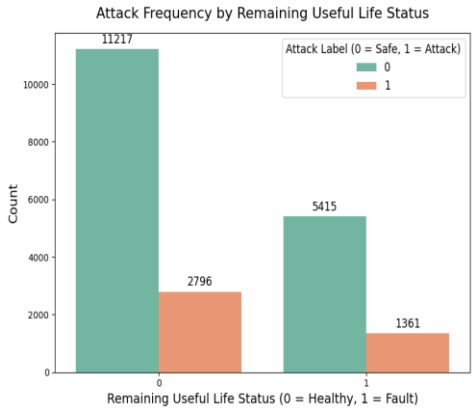

Figure 1. Distribution of Samples Across Attack Categories

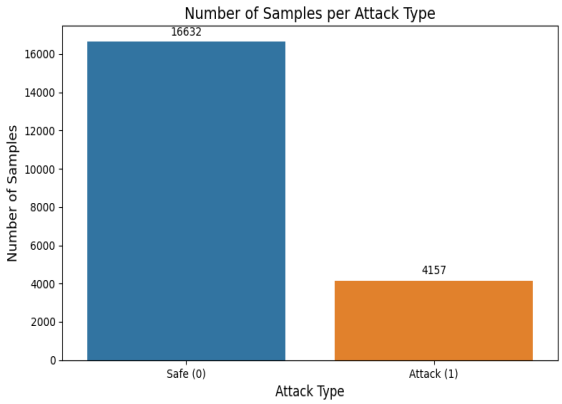

Figure 2. Attack Distribution Across Engine Health Conditions

Figure 1, represents the sample distribution of the two types of attacks and it is clear that the number of normal observations is much more than the adversarial cases. Most of the data is safe with a relatively smaller ratio being related to manipulated data. This skew illustrates the need of using data balancing methods in preprocessing to avoid favouring the majority group in model training.

As shown in figure 2, the adversarial activity of a healthy and faulty engine is different. In the healthy samples, the number of safe records and the number of manipulated records is greater than, in faulty samples, which implies the presence of the adversarial interference in both operating conditions with varying frequencies. This distribution makes the modelling methods that are useful in different engine health conditions important because adversarial behaviour may arise irrespective of the mechanical condition.

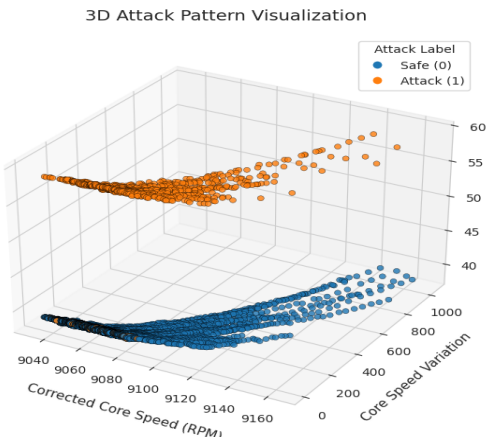

Figure 3. Three-Dimensional Visualization of Attack Patterns

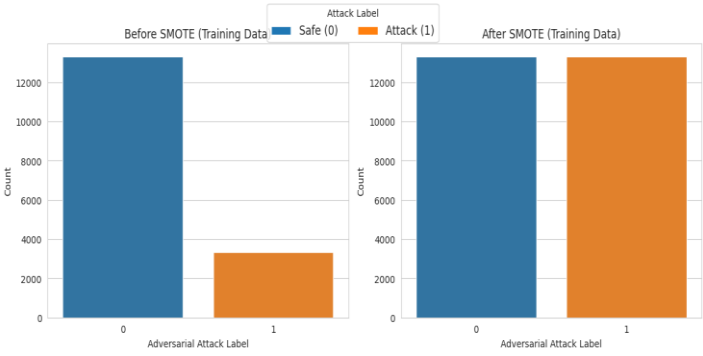

Figure 4. Class Distribution Before and After SMOTE Application

Figure 3 demonstrates a three-dimensional split between safe and adversarial samples, adversarial points are placed in a separate area, which is evidence of obvious changes in sensor relationships in conditions of manipulation. This visual difference shows that adversarial samples change the interaction between features in a manner that can be learnt by models.

The image below (Figure 4) shows the distribution of classes before and after the application of SMOTE. The dataset is characterized by safe samples, however, once SMOTE oversampling is used, the adversarial group becomes even with the majority group, making the learning process fairer and avoiding bias in the model.

3.5 FT-Transformer Architecture and Model Rationale

The FT-Transformer is based on a token representation in which every numeric feature is transformed to a learnable embedding, which allows the architecture to process tabular inputs in a structured sequence form. Each feature token is embedded into a column to maintain identity of features, and a classification token is also included to combine global information of the feature sequence. These tokens are processed by the model in several transformer encoder layers, which use multi-head self-attention to model inter-feature dependencies. The design enables the network to study the impact of each variable on other instead of depending on the contribution of each feature.

FT-Transformer has a powerful technical benefit in adversarial detection, which is provided by the attention mechanism. The adversarial interference tends to interfere with relationship between parameters of the engine and maintain the value of individual features in a plausible domain. To learn and handle such deviations, the multi-head attention operation dynamically adjusts attention weights between features enabling the model to learn abnormal patterns of interaction even when the perturbations are small. Additional elements of the network such as layer normalization, feed-forward sublayers, and residual connections make the process of learning more stable, which makes it more resistant to distributional shifts caused by manipulated inputs.

This architecture focuses on the drawbacks of the traditional machine learning and deep learning architectures where the feature hierarchies or decision boundaries can often be subject to attack through targeted perturbations. Conversely, the attention-based design re-calculates feature relevance to each input, making it less vulnerable to adversarial noise and more predictable to behave in altered environments. The transformer architecture does not also pay the computational cost of recurrent or convolutional networks, which is a highly scalable solution to complex sensor-based tasks, and adversarial adjustments can introduce distortions to the signatures of functions.

3.6 Evaluation Metrics

In order to determine reliably model performance with class imbalance, this paper compares all the models with macro-averaged precision, recall, and F1-score besides accuracy. Since the dataset has much more clean samples as compared to adversarial samples, the standard accuracy could seem high, even when a model is performing poorly on the minority class. The macro averaging calculates each measure separately of the two classes and then averages them without weighting such that the minority (attack) and the majority (safe) class contribute equally. This methodology gives a more balanced and realistic estimate of the performance of adversarial detection especially in the context of assessing the ability of each model to identify manipulated inputs.

4 Design Specification

This chapter presents the design requirements of the adversarial attack is stability, modularity, and consistent performance in the sense that every stage adds value to the process of detecting the adversarial influence in the engine sensor settings.

4.1 System Architecture Overview

Figure 5 shows the system architecture of the system, which is organized into a series of interrelated layers that control the entire process of converting raw engine sensor data to final model evaluation. The architecture starts with the data collection layer where sensor values of turbofan engine and their corresponding labels of either normal or adversarial conditions are obtained. These crude inputs are authenticated to be structurally consistent, complete and correct to make sure that all the measurements are appropriate to the downstream analysis. After validation, the data is fed into the pre-processing layer which forms a crucial base of all modelling methods. This layer will do a number of transformations such as data cleaning, relevant variables encoding, exploratory data analysis, redundant features removal, class rebalancing via SMOTE, numerical scaling, and dividing the dataset into two parts, training and testing. All these steps will produce a stable and consistent dataset that can be used in the study with each type of models.

After pre-processing, the architecture goes to modelling layer, in which three modelling branches run simultaneously. Traditional machine learning models, which are the first branch, are Logistic Regression, Random Forest, and LightGBM, all of which are based on a statistical or tree-based decision mechanism. The second branch is comprised of deep learning models, such as Multi-Layer Perceptron and a custom convolution-based neural network which is developed to learn non-linear feature patterns using trainable neural architectures. The third branch is associated with the FT-Transformer model, which breaks down the individual features and uses multi-head self-attention, residual connections, normalization, and feed-forward transformations to capture fine-grained relationship among sensor variables. These branches are different in their internal mechanisms although they have the same pre-processed input to compare them fairly.

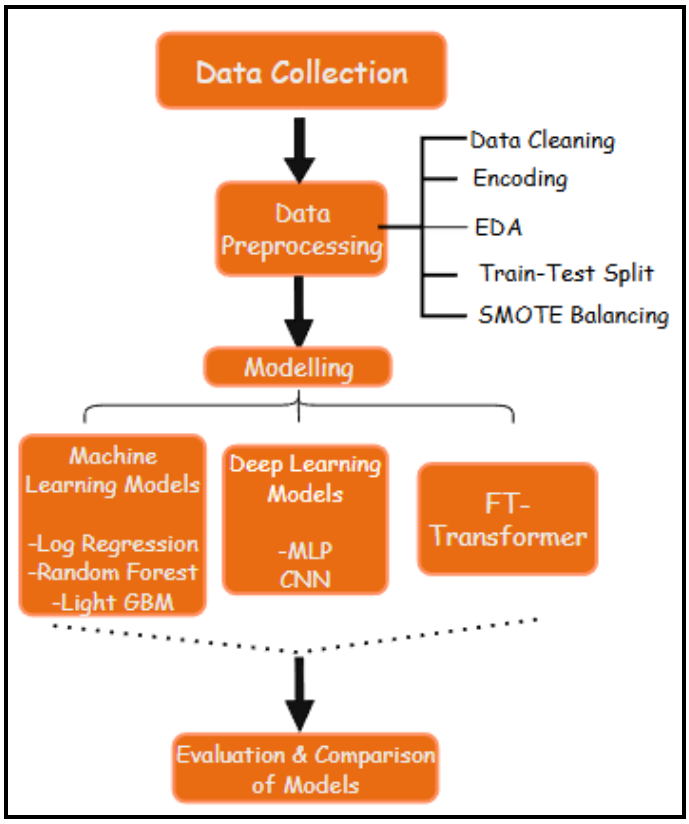

Figure 5. End-to-End Methodology Pipeline

The last tier of the architecture is on the evaluation and comparison. After training all the models, their accuracy, macro-averaged precision, recall and F1-score are used to evaluate their predictions on unseen data to encompass both clean and adversarial classes. Confusion matrices are analysed in order to determine the error tendencies and have a better understanding of detection reliability. The architecture, through combining data collection, pre-processing, parallel modelling, and comparative evaluation into one structured pipeline, ensures that there is coherence, reproducibility, and clarity in the evaluation of methods of adversarial attack detection.

4.2 Workflow and Pipeline

The sequence of workflow directing the system takes a simplified form whereby the sensor dataset, that includes both normal and adversarial samples, is loaded into the environment to be analysed. Once ingested, the data is pre-processed, and it is verified against inconsistencies, cleaned, balanced with SMOTE, and deprived of redundant features as well as standardized to produce a stable and uniform input. After preparing, the dataset goes through the modelling phase, either through first being turned into tokenized embeddings with the FT-Transformer or simply fed directly into other models to train to find differences between clean and manipulated readings. The FT-Transformer processes operate by using attention-based encoder layers to imply abnormal interaction patterns brought about by adversarial perturbations. The last process in the workflow is assessment of unseen data, in

which accuracy, macro precision, recall, and F1-score are computed to evaluate the detection performance and to identify whether each sample represents normal or adversarial behaviour. This end-to-end pipeline provides a stable and reliable system of adversarial attack detection in sensor-driven settings.

5 Implementation

The implementation phase is concerned with the last stage of development and optimization of all machine learning, deep learning, and transformer-based models utilized to detect adversarial attacks. Both models have been trained in a controlled setting where sensor data, having undergone all preprocessing and feature preparation procedures, was fed into the training phase which was assisted by systematic hyperparameter tuning.

5.1 Machine Learning Models Implementation

The machine learning models applied in the current study were conducted by means of structured hyperparameter tuning procedures that were aimed at finding stable configurations of adversarial detection. The logistic Regression was optimized with six values of the regularization constant c (0.001, 0.1, 0.1, 1, 10, 100) which were evaluated using five-fold cross-validation and full data training and kept c = 0.001 as the most stable setting. Figure 6 was used to tune the Random Forest classifier to 50, 100, 150, 200, and 250 estimators configuration with each configuration being evaluated using cross-validation and independent retraining and the finalized configuration was 150 estimators. LightGBM was fine-tuned by testing 5 learning rates (0.001, 0.01, 0.05, 0.1, and 0.5) at a fixed number of 100 estimators, and cross-validation, as well as full retraining, proved that 0.5 was the most stable learning rate. In all models, tuning was done by systematically exploring parameters based on five-fold cross-validation and secondary stability checks on the full training data, so that each identified configuration corresponded to reliable behaviour on the adversarial detection task.

5.2 Deep Learning Models Implementation

Structured neural architecture was used to implement the deep learning models that included controlled regularization, adaptive optimization, and training behaviour monitoring. The Multi-Layer Perceptron (MLP) was designed in a sequential architecture with hidden layers of 128 and 64 with ReLU activation with a dropout of 0.2 and a sigmoid output unit to make binary classification. It was also trained with the Adam optimizer with a learning rate of 0.001, binary cross-entropy loss, a batch size of 256, a validation split of 20 percent and an early stopping criterion that tracked validation loss with a patience level of epoch. The custom deep neural network had a deeper architecture with three hidden layers with 256, 128, and 64 units with the first two layers having L2 regularization with a strength of 1e-4, batch normalization, and dropout rates of 0.3 and 0.2 respectively. The Adam optimizer with learning rate 0.001, binary cross-entropy loss, early stopping of training on validation accuracy with patience of epochs, and a learning-rate reduction mechanism reducing the learning rate by a factor of 0.5 when validation loss stalled and with a minimum possible

learning-rate of 1e-6 were used to manage the training. Each of the two models was trained to a maximum of 200 to 300 epochs, and the deep learning implementation was finalized with the help of the properly chosen layer sizes, regularization measures, and dynamic training controls.

5.3 FT-Transformer: Parameter Design

To attain its high performance, the FT-Transformer transformed every sensor feature into a learnable 128-dimensional token and augmented it with a column embedding that maintains the identity of its features to enable the model to process tabular inputs like sequential data. These tokens, in addition to a learnable CLS token, are encoded by four transformer encoder layers with eight attention heads, allowing the network to learn more nuanced and complex interactions, which adversarial perturbations bring about between engine parameters. The architecture employs GELU activations, residual links, and layer normalization to ensure stable and efficient learning and dynamically re-estimates the importance of features to each input sample. The output of the transformer is converted by a fully connected classification head into a binary prediction using a sigmoid function. The AdamW optimizer with a low learning rate of 5e-5 and cosine annealing to smoothly change the learning-rate schedule is used to optimize the training, and early stopping is used to prevent overfitting of the model. A combination of these mechanisms enables the FT-Transformer to identify fine-grained adversarial changes in sensor behaviour, which leads to high accuracy, recall, and F1-score than other baseline models.

5.4 Tools and Libraries Used

The combination of Python-based tools and machine learning libraries made this study possible and facilitated the necessary data preprocessing, model development, training, and evaluation. Python 3.9 was the main programming environment because it has a large number of scientific computing packages. NumPy and Pandas libraries were employed to perform numerical operations and manipulate data whereas Matplotlib and Seaborn were used to create exploratory data visualization that showed patterns of distributions and the relationship between features. Scikit-learn offered the implementations of classic machine learning models, preprocessing tools, hyperparameter optimization, and evaluation measures. Imbalanced-learn (SMOTE) was applied to equalize the classes prior to training the models. TensorFlow and Keras have helped in creating deep learning models, which provide easy APIs to build and optimize neural architectures.

6 Results and Discussion

In this chapter, the experimental findings were provided with the FT-Transformer model and the baseline classifiers that were compared to it. The effectiveness of each model is determined based on the measures of accuracy, classification reports and confusion matrices to determine that the model can distinguish between adversarial samples and non-adversarial samples of the engine sensor dataset. The results of such measurements give an indication of

the strong and weak sides of the respective models and indicate the relative strength of the FT-Transformer to deal with adversarial influenced sensor data.

Table 2. Model Performances Results

Model Name	Accuracy	Macro Averaged Precision	Macro Averaged Recall	Macro Averaged F1 Score
Logistic Regression	0.70	0.58	0.61	0.59
Random Forest	0.84	0.91	0.60	0.62
Light GBM	0.61	0.51	0.52	0.50
Multi-Layer Perceptron	0.84	0.92	0.60	0.62
Custom Neural Network	0.84	0.92	0.60	0.62
FT-Transformer	**0.93**	**0.95**	**0.82**	**0.87**

The table 2 results, give us a relative analysis of all the models and clearly indicate the distinct differences in the ability of the models to identify adversarial samples. The performance of the traditional machine learning and deep learning models is moderate with a considerable decrease in recall meaning that it is not easy to uncover manipulated data consistently. FT-Transformer is superior to all baselines in terms of accuracy, precision, recall, and F1-score, which shows that it is more robust and has a better ability to interact with features in adversarial conditions.

6.1 Performance of Machine Learning Models

6.1.1 Logistic Regression

The optimal value of C is found, the Logistic Regression model is retrained with this hyperparameter and its performance is assessed with classification metrics. The model has a test accuracy of 0.70 with a macro precision of 0.58, macro recall of 0.61 and macro F1-score of 0.59 indicating a moderate ability to discriminate. As indicated by the classification report, the model has a good performance on the majority (safe) class but its recall on the adversarial type is low which means it cannot recognize the manipulated samples with a constant success.

6.1.2 Random Forest

The model has a test accuracy of 0.84, macro precision of 0.91, macro recall of 0.60 and macro F1-score of 0.62. According to the classification report, the model is doing extremely well on most of the classes, with the highest recall of almost perfect detection of clean samples, but it has a hard time classifying the adversarial samples, as the recall of the attack class is 0.20. This difference indicates that whereas the ensemble generalizes well with

general patterns, adversarial perturbations break decision path in individual trees, preventing the model to generalize to the minority class.

6.1.3 LightGBM

Upon tuning, the best model is tested on the test data and the accuracy is 0.6056, which is significantly lower than other the baseline models. According to the classification report, LightGBM has a moderate performance: it works on the majority (clean) class with a precision of 0.81 and a recall of 0.67, but it has a very low performance on the adversarial samples, having 0.21 precision and 0.36 recall on the minority class. This discrepancy indicates that although the gradient-boosting mechanism is very powerful, the leaf-wise split strategy of the model can be unstable in cases where adversarial perturbations cause the feature boundaries to shift in a subtle way.

6.2 Performance of Deep Learning Models

6.2.1 Multi-Layer Perceptron (MLP)

The model has a total accuracy of 0.84 however a closer examination of the classification report shows that there is a significant disparity in performance between classes. The MLP is almost perfect in predicting clean samples with a precision of 0.83 and recall of 1.00, showing that the network is highly skewed towards the majority class. The model however has a recall of only 0.20 on adversarial samples indicating that it is not effective in recognizing manipulated examples even with high precision. This difference indicates that although the MLP can learn non-linear patterns, adversarial perturbations destabilize internal activation pathways resulting in low sensitivity to the minority class. This imbalance is supported by the confusion matrix, which indicates that there were many false negatives with adversarial attacks.

6.2.2 Custom Neural Network

The last analysis demonstrates that the custom neural network has a precision of 0.84, which is equal to the performance of the MLP but demonstrates identical imbalance in the detection abilities of the classes. The model is highly precise and recalls all clean samples with a high recall rate and is effective in identifying most of the classes without misclassifying. Nevertheless, its recall on adversarial samples is only 0.20, which means that it is challenging to identify perturbed inputs even with deeper layers and regularization, as well as dropout. This implies that despite the architecture being very effective in capturing non-linear feature structures, adversarial noise continues to perturb the activation paths of the network, and does not allow the manipulated class to be detected consistently.

6.3 Performance of the FT-Transformer Model

The FT-Transformer showed the best performance among all the models considered and obtained an accuracy of 0.93 with a high level of precision, recall, and F1-score. The model was able to detect the subtle deviations caused by adversarial manipulation because of its attention-driven architecture, which made it analyse dynamic relationships between engine parameters. The feature representation in the form of tokens enabled every variable to provide contextual contribution to the classification process, which enabled the model to be more robust to the effects of perturbation. In comparison to linear, tree-based, or dense neural models, the FT-Transformer re-computed the relevance of features per input instance, enhancing its capability to discriminate clean samples and adversarial ones when the perturbations were similar to natural patterns. Such a uniform performance in metrics demonstrates that the model is suitable in the case of adversarial detection in sensor-driven settings.

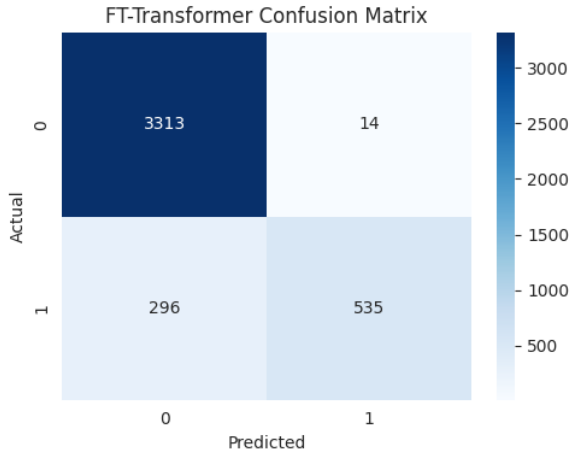

Figure 6. FT-Transformer Confusion Matrix

Figure 6, presents the confusion matrix of the FT-Transformer, which has a good performance of identifying clean samples correctly with only a few false positives. Even though there were some adversarial examples that were falsely classified as clean, the model demonstrated a significant percentage of the manipulated samples and this indicates that it is able to identify the patterns of perturbations. All in all, predictions distribution proves to have a high degree of reliability in identifying the normal and adversarial engine sensor readings.

The FT-Transformer was more accurate because of its attention-based processing that enables the model to dynamically assess the relationship among features in regard to each input instance. Adversarial perturbation tends to change the interaction patterns of engine parameters and does not significantly change the individual ones. The multi-head self-attention mechanism accounts such subtle deviations by placing the importance of each feature based on the context so that the model is able to identify the inconsistencies that might be missed by traditional and deep learning models. This calibration capability of

feature relevance in inference contributes to the overall consistency with which adversarial patterns are recognized, which is a direct contribution to its high performance.

The other benefit that has an impact on the accuracy of the model is that the model is represented in the form of tokens, with each feature changed into an embedding supplemented with column-level identifiers. This structure retains feature identity and permits the transformer encoder to manipulate variables in a semantic sequence-like form. Additional stabilization measures include residual connections, layer normalization, and feed-forward sublayers which mitigate the effects of noises and distribution changes that can be caused by adversarial manipulation. This leads to the FT-Transformer having strong decision boundaries and better recall and F1-score than other models, demonstrating its usefulness in complicated sensor settings.

6.4 Comparison of Models

Comparison of all the models brings out significant differences in the capability of all the models in the detection of adversarial manipulation in the engine sensor data. LightGBM and Logistic Regression, which are traditional models of machine learning, performed the worst because the use of linear or tree-based decision boundaries meant that these models were more vulnerable to distributional shifts brought about by adversarial perturbations. Random Forest was more precise because of its ensemble form but weak recall, which means that the method is not always able to detect manipulated samples under different conditions. Deep learning models, such as Multi-Layer Perceptron and the Custom Neural Network, were more accurate and precise because they learn non-linear features representation. Nonetheless, the two networks were moderate in recall, which indicates the impact of adversarial changes on their internal activations. These models could be used to classify clean samples but were less reliable in the face of perturbed inputs that resemble natural patterns.

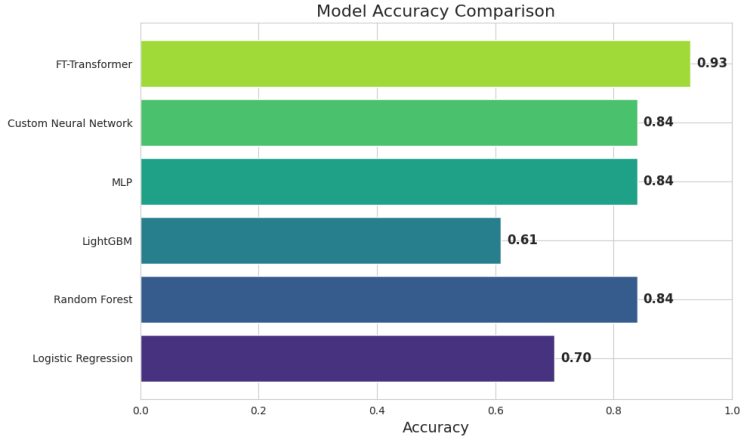

Figure 7. Model Accuracy Comparison

Figure 7, which compares the accuracy of all the models that were evaluated, demonstrates the obvious performance differences among approaches. Moderate and high accuracy is attained by traditional, neural models, similar results were attained by the Random Forest,

MLP, and the Custom Neural Network. Logistic Regression and LightGBM are doing relatively worse. The FT-Transformer is the most effective model to be used in the detection of adversarial manipulation in sensor data because it has the highest accuracy of 0.93.

On the contrary, the FT-Transformer achieved high accuracy, precision, recall, and F1-score compared to all baseline models. Its architecture based on attention enabled it to learn finer level of interactions between features, which made it more skillful in detecting inconsistencies due to adversarial manipulation. The overall superiority of metrics suggests that transformer-based processing of features offers a more robust and flexible architecture of adversarial detection than traditional machine learning and deep learning methods.

7 Conclusion

This paper explored the topic of adversarial manipulation on engine sensor data by evaluating the performance of various machine learning, deep learning, and transformer-based models. It was shown through the analysis that the classical models, despite having general operational patterns, had low resilience to perturbed inputs. The logistic regression, random forest, lightgbm and the two deep neural networks demonstrated different levels of accuracy yet they were uniformly poor at recalling adversarial samples, which is indicative of their inability to detect subtle deviations that are being introduced to sensor measurements. These results support the current literature that cites fixed-weight architecture and fixed feature representations as some of the most important factors leading to vulnerability in adversarial conditions.

Conversely, the FT-Transformer provided much higher results in all the measures of evaluation. Its attention-based architecture made it easier to model feature interaction and it was able to identify irregularities in sensor relationships which are not apparent in linear or densely connected models. The F1 and the high recall scores of the FT-Transformer show a stronger ability to distinguish between clean and manipulated readings, which proves that it is appropriate in the context of adversarial detection in sensor-based industrial settings. The comparative evaluation of all baselines gives a good indication that feature tokenization on attention can improve the system to stay stable and reliable when affected by adversarial influence.

7.1 Future Work

The extension of the transformer-based framework to larger and more heterogeneous sensor data can be investigated in the future to examine how well it can adapt to various industrial settings. The addition of real time adversarial detection mechanisms and online learning approaches can also lead to an increased responsiveness of the model to changing attack patterns. Further research is possible on hybrid systems integrating attention with generative or self-supervised learning to enhance robustness in conditions of limited or noisy data. The connectivity with edge-deployed systems and computational efficiency optimization are also helpful directions of the development of practical deployment

References

Zhao, J., Xie, L., Gu, S., Qin, Z., Zhang, Y., Wang, Z. and Hu, Y., 2025. Universal attention guided adversarial defense using feature pyramid and non-local mechanisms. *Scientific Reports*, *15*(1), p.5237.

Xie, Y., Wang, J., Xie, S. and Chen, X., 2023. Adversarial training-based deep layer-wise probabilistic network for enhancing soft sensor modeling of industrial processes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *54*(2), pp.972-984.

Guo, R., Chen, Q., Liu, H. and Wang, W., 2024. Adversarial robustness enhancement for deep learning-based soft sensors: An adversarial training strategy using historical gradients and domain adaptation. *Sensors*, *24*(12), p.3909.

Vitorino, J., Praça, I. and Maia, E., 2023. Towards adversarial realism and robust learning for IoT intrusion detection and classification. *Annals of Telecommunications*, *78*(7), pp.401-412.

Shihab, M.A., Marhoon, H.A., Ahmed, S.R., Radhi, A.D. and Sekhar, R., 2024. Towards resilient machine learning models: Addressing adversarial attacks in wireless sensor network. *Journal of Robotics and Control (JRC)*, *5*(5), pp.1599-1617.

Owezarski, P., 2023, May. Investigating adversarial attacks against Random Forest-based network attack detection systems. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium* (pp. 1-6). IEEE.

Gaber, T., Ali, T., Nicho, M. and Torky, M., 2025. Robust Attacks Detection Model for Internet of Flying Things based on Generative Adversarial Network (GAN) and Adversarial Training. *IEEE Internet of Things Journal*.

Haroon, M.S. and Ali, H.M., 2022. Adversarial Training Against Adversarial Attacks for Machine Learning-Based Intrusion Detection Systems. *Computers, Materials & Continua*, *73*(2).

Sauka, K., Shin, G.Y., Kim, D.W. and Han, M.M., 2022. Adversarial robust and explainable network intrusion detection systems based on deep learning. *Applied Sciences*, *12*(13), p.6451.

Albattah, A. and Rassam, M.A., 2023. Detection of adversarial attacks against the hybrid convolutional long short-term memory deep learning technique for healthcare monitoring applications. *Applied Sciences*, *13*(11), p.6807.

Li, J., Yang, Y., Sun, J.S., Tomsovic, K. and Qi, H., 2023, July. Towards adversarial-resilient deep neural networks for false data injection attack detection in power grids. In *2023 32nd*

International Conference on Computer Communications and Networks (ICCCN) (pp. 1-10). IEEE.

Anasuri, S., 2022. Adversarial Attacks and Defenses in Deep Neural Networks. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), pp.77-85.

Mustapha, A., Khatoun, R., Zeadally, S., Chbib, F., Fadlallah, A., Fahs, W. and El Attar, A., 2023. Detecting DDoS attacks using adversarial neural network. *Computers & Security*, *127*, p.103117.

Alzaidy, S. and Binsalleeh, H., 2024. Adversarial attacks with defense mechanisms on convolutional neural networks and recurrent neural networks for malware classification. *Applied Sciences*, *14*(4), p.1673.

Huang, R. and Li, Y., 2022. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. *IEEE Transactions on Smart Grid*, *14*(3), pp.2367-2376.

Barik, K., Misra, S. and Fernandez-Sanz, L., 2024. Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network. *International Journal of Information Security*, *23*(3), pp.2353-2376.

Alshahrani, E., Alghazzawi, D., Alotaibi, R. and Rabie, O., 2022. Adversarial attacks against supervised machine learning based network intrusion detection systems. *Plos one*, *17*(10), p.e0275971.

Dai H, Wu S, Huang J, Jian Z, Zhu Y, Hu H, Chen Z. FT-Transformer: Resilient and Reliable Transformer with End-to-End Fault Tolerant Attention. arXiv preprint arXiv:2504.02211. 2025 Apr 3.