

A Deep Learning Framework for Detecting Cyber Vulnerabilities in IOT-Based Data Exchange Systems

Ezeibeonu Ozioma Stephanie

Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli,
Anambra State

ezeibeanustephanie@gmail.com

ABSTRACT

The increased use of the Internet of Things (IoT) has enabled improved data exchanges that span various fields, such as healthcare, industrial systems, smart cities, and transportation. However, the growth of the IoT paradigm has presented greater challenges in securing data exchanges due to the increased attack surfaces created by connected devices. Threats such as Distributed Denial-of-Service (DDoS), malware, botnets, brute-force, and web intrusions can target IoT-based systems, thus compromising the confidentiality, integrity, and availability of data. Current cybersecurity solutions are inadequate in detecting advanced persistent threats due to their limited ability to understand the temporal relationships between network traffic features. This work proposed a Temporal Fusion Transformer (TFT)-based approach to identifying cyber vulnerabilities in IoT-based data exchange systems. The publicly available Edge-IIoTset dataset was used for benchmarking purposes. The data was preprocessed, including cleaning, removing duplicates, handling missing values, encoding, Min-Max normalization, feature selection, balancing class distributions using Synthetic Minority Oversampling Technique (SMOTE), and sequential data preparation. The proposed TFT was trained and validated using the Adam optimizer and categorical cross-entropy loss, with a 70:15:15 training/validation/testing split and early stopping. The model achieved a testing accuracy of 93.4%, precision of 93.8%, recall of 92.6%, an F1-score of 92.8%, 96.8% ROC-AUC, and a Matthew's correlation coefficient (MCC) of 0.912. Additionally, the framework registered a false positive of 4.2%, 34.5ms detection latency, and 0.52ms inference time per batch. TFT surpassed traditional and state-of-the-art deep learning counterparts in detecting different categories of IoT-based cyberattacks. The work concluded that the proposed framework could offer a reliable and scalable solution to securing IoT-based data exchange systems.

KEYWORDS: Internet of Things; Cyber Vulnerability Detection; Temporal Fusion Transformer; Deep Learning; Edge-IIoTset

I. INTRODUCTION

Internet of Things (IoT) has become one of the most disruptive concepts in the current digital era by enabling connected devices to communicate and exchange data over the internet. IoT-based data exchange systems use sensors, actuators, gateways, cloud, and other components to offer smart solutions in health-care, home appliances, industrial systems, farming, transportation, and smart cities (Agiillo et al., 2021). The increased use of IoT has created significant economic value while at the same time introducing vulnerabilities that undermine the confidentiality, integrity, and availability of data in various application areas (Fatorachian and Kazemi, 2021). Throughout this manuscript, the following terminology is used with the meanings as defined below. A cyber vulnerability is defined as a weakness in an IoT device, protocol or a data exchange mechanism that can be exploited by an adversary. In turn, a cyber threat or attack refers to actual exploitation that occurs in the system. Finally, intrusion detection is the process of detecting attacks within traffic data. With this basis, our

framework allows for detecting cyber vulnerabilities that manifest themselves as anomalous temporal patterns in the traffic data, thus providing a general intrusion detection mechanism. In particular, note that in this manuscript, vulnerability detection is different from intrusion detection; the former refers to finding weaknesses before they are actually exploited, while the latter analyses traffic data in order to classify it as normal or malicious (Pourrahmani et al., 2023).

There are various sources of cyber vulnerabilities in IoT-based data exchange systems, including insecure communication protocols, weak authentication and access control, lack of encryption, firmware weaknesses, software design flaws, and the use of resource-constrained devices (Williams et al., 2022). Attackers can target these weaknesses and use cyber vulnerabilities such as DDoS, malware, botnets, spoofing, man-in-the-middle, data manipulation, and eavesdropping to access data, disrupt services and take over devices. Traditional cybersecurity solutions such as signature-based intrusion detection systems and vulnerability scanners are ineffective in detecting zero-day attacks and sophisticated persistent threats due to their limited ability to capture temporal relationships between network traffic features (Bin Hulayyil et al., 2023; Chowdhury et al., 2024). Recently, there has been an emerging need to secure IoT-based data exchange systems through artificial intelligence-driven approaches. The deep learning (DL) based techniques excel at representing hierarchically structured data derived from network and system logs, which facilitates the identification of different types of cyber vulnerabilities (Raavi, 2025). Indeed, it has been reported that convolutional, recurrent, and transformer-based neural networks, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Autoencoder, Temporal Fusion Transformer (TFT), and other architectures, were applied for intrusion detection, network anomaly detection, and traffic classification to mine spatial and temporal patterns of network traffic (Aldhaheri et al., 2024).

Recently, the transformer-based encoder-decoder architecture was found to be capable of encoding long-range dependencies in a sequential data better than the recurrent layer, which explains its recent adoption in the network traffic analysis task (Zhang et al., 2024). On the other hand, the explainable AI (XAI) methods, such as attention visualization, SHAP, and counterfactual explanations, have been utilized to enhance the security of the IoT-based data exchange systems, as they enable auditing of the DL model decisions, thus complying with the regulatory requirements of algorithmic transparency and accountability (Kumar & Singh, 2025). Notably, federated learning was advanced as a privacy-preserving threat detection framework for IoT networks, which allows collaborative model training across edge devices while maintaining the confidentiality of the data (Alwahedi et al., 2024). All in all, the existing research has not yet converged on applying temporal fusion networks for the pre-attack vulnerability discovery task, which motivates the current study. Moreover, most prior works focused on detecting attacks once they occurred instead of analyzing network traffic to uncover potential cyber vulnerabilities that could be used to launch attacks. Thus, there is a need for an efficient and reliable deep learning framework that can help secure IoT-based data exchanges by detecting cyber vulnerabilities and weaknesses in connected devices, networks, and systems before they can be exploited. This work proposed a deep learning framework that leverages Temporal Fusion Transformers to detect potential cyber vulnerabilities in IoT-based data exchange systems. The next section discusses the methodology used in this work.

II. METHODOLOGY

This work made use of a deep learning approach to detect potential cyber vulnerabilities in IoT-based data exchange systems. Benchmark datasets with information about normal and malicious network traffic were collected and combined where appropriate for this study. The data was pre-processed and partitioned into training, validation, and testing sets. A Temporal Fusion Transformer was applied to detect network anomalies and cyber vulnerabilities. The performance of the proposed approach was then evaluated using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, MCC, false positive rate, detection latency, and computational cost. This framework was compared to conventional machine learning and deep learning methods to demonstrate its efficacy in detecting cyber vulnerabilities in IoT-based data exchange systems.

A. Data Acquisition (Edge-IIoTset)

The data used in this work was obtained from the Edge-IIoT set of the Industrial IoT and consists of network traffic data generated by various IoT devices, such as sensors, actuators, cameras, and other connected components working in tandem with cloud and edge servers. The data was generated by simulating various cyberattack scenarios, including but not limited to DDoS, Denial of Service (DoS), brute force, information gathering, web intrusions, malware campaigns, and botnet attacks. The dataset consists of multiple training examples with rich network traffic features and labels indicating either normal or attack traffic. For this work, the dataset was downloaded from the official public repository and used to train, validate, and test the proposed framework. Some preprocessing steps needed to be carried out before the data could be used for training the TFT model. Initially, the data was analyzed to remove irrelevant features, duplicates, and missing features. Numerical features with missing values were estimated using the median algorithm while the categorical values were replaced using the label encoding technique. The data was then normalized using the Min-Max normalization function to ensure that all features have equal weights in the model. Furthermore, feature selection and dimensionality reduction were carried out to remove highly correlated features with little effect on the performance of the model.

The network traffic data was then divided into sequential time windows and used to train the Temporal Fusion Transformer to detect cyber vulnerabilities in IoT-based data exchange systems. Furthermore, SMOTE was applied to the training data to balance the class distribution and improve the model's performance in detecting rare attack classes. The data was then split into training (70%), validation (15%), and testing (15%) sets. In order to verify the generalization ability and cross-domain robustness, an extra benchmark dataset, namely UNSW-NB15, is introduced in our experiments. UNSW-NB15 mainly consists of network traffic data, which differs greatly from the IoT traffic data in terms of feature distribution and attack patterns. The same preprocessing procedure (cleaning, encoding, normalization, SMOTE, and sequential windowing) are conducted on the new dataset. We then perform two sets of experiments on this dataset: (i) direct cross-domain test, in which the TFT model only pre-trained on Edge-IIoTset is directly deployed on the new domain without any finetuning; (ii) transfer learning with few-shot finetuning in which 10% of the training data are used to finetune the pre-trained model for another 5 epochs. By performing these two sets of experiments, we aim to decouple the contribution of the learned temporal representations and the influence of domain specific feature distribution.

B. Data Preprocessing

The data gathered from the Edge-IIoTset dataset needed to be pre-processed and prepared to ensure it was clean, complete, and ready for use in training the Temporal Fusion Transformer. Initially, the data was analyzed to identify and remove irrelevant features, duplicates, and missing features. Numerical features with missing values were estimated using the median algorithm while the categorical values were replaced using the label encoding technique. The data was then normalized using the Min-Max normalization function to ensure that all features have equal weights in the model.

C. The Proposed Temporal Fusion Transformer (TFT) Model

The Temporal Fusion Transformer (TFT) is a state-of-the-art architecture capable of capturing complex temporal features in multivariate sequential data with high accuracy and model interpretability. Unlike the conventional Long Short-Term Memory (LSTM), the TFT model has an enhanced ability to capture both short-term and long-term temporal patterns using combination of recurrent and attention mechanisms, as well as variable selection networks (LSTM Encoder and Decoder). This makes the TFT ideal for use in detecting cyber vulnerabilities in IoT-based data exchange systems, since network traffic has short- and long-term temporal patterns that need to be captured and used to detect possible cyber problems.

The proposed TFT model for IoT-based data exchange vulnerability detection consists of several modules, namely; the Variable Selection Network (VSN), Multiple Attention Gates (MAG), LSTM Encoder and Decoder, and the Temporal Fusion Network (TFN), as shown in Figure 1. Each of these components plays a significant role in ensuring the model can capture relevant features in IoT network traffic data and use them to detect vulnerabilities in data exchange systems. The VSN component helps to identify the most relevant features that should be used in the model by assigning weights to different input features. The weighted features are then passed through the embedding layer to generate higher-order representations that help in capturing complex patterns in the input data. Furthermore, the MAG mechanism allows the model to selectively focus on specific aspects of the encoded input data that are relevant to the task at hand, such as detecting anomalies or vulnerabilities in the IoT network or data exchange systems. On the other hand, the LSTM encoder and decoder play a critical role in capturing short-term temporal patterns in the sequential input data that the model receives. The encoder processes the input sequence and generates a hidden state that is passed to the decoder, where it is further processed to generate an output sequence. This output sequence is then combined with the contextual information generated by the TFN to make predictions about the network traffic data.

To capture long-range interactions between communication events, the encoder outputs are transformed using an interpretable multi-head self-attention mechanism, where attention parameters learn to focus on the most informative context around each time-step and feature to gain better understanding on the cyber vulnerability prediction. Gating and residual connections are applied to suppress irrelevant information, facilitate information flows, and stabilize the training of the TFT. Contextual representations from the attention layer are then processed using fully connected layers and the output is fed into a Softmax classifier, which labels each communication sequence as normal or of any type that corresponds to the cyberattack classes. The proposed TFT will be trained using the Adam optimizer and a categorical cross-entropy loss objective, where early stopping and dropout will be applied to avoid overfitting during the training. Such hyperparameters as learning rate, batch size, attention heads, the number of hidden units, dropout levels, and sequence length will be tuned to optimize the performance on the validation set.

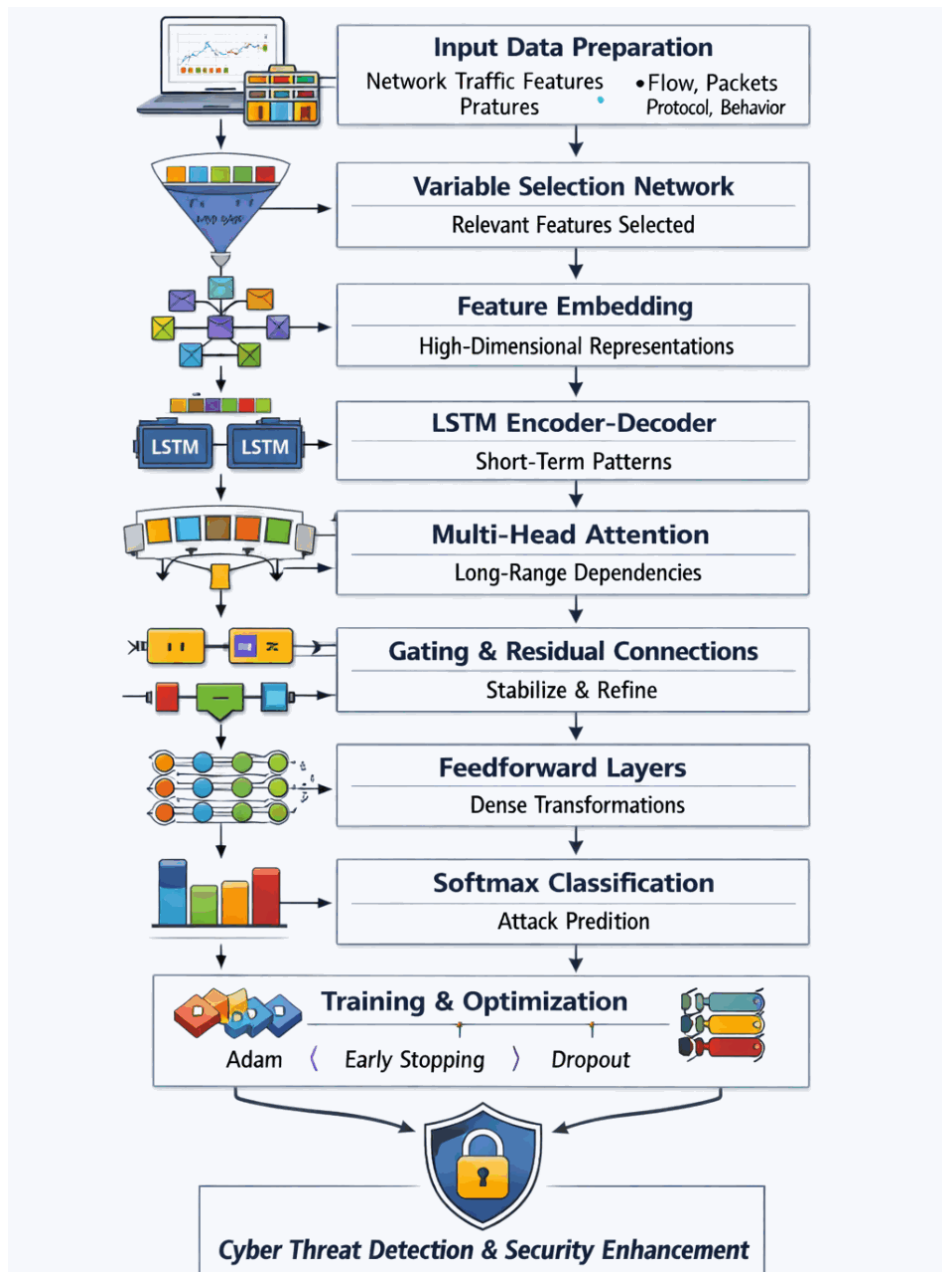


Figure 1: The Architectural Flow Diagram of the Proposed TFT Model

As a result, we expect that the trained TFT will be able to learn the input sequential communication patterns, which can help detect the potential cyber vulnerabilities to increase the overall security and trustworthiness of IoT-based data exchange systems.

Table 1: Architecture Parameters of the Proposed TFT Model

Parameter	Specification
Model Architecture	Temporal Fusion Transformer (TFT)
Input Data	Preprocessed sequential network traffic from Edge-IIoTset
Input Features	Selected network flow and communication features
Sequence Length	64 time steps
Variable Selection Network	Enabled
Embedding Dimension	128
LSTM Hidden Units	128
LSTM Layers	2
Multi-Head Attention Heads	4
Dropout Rate	0.20
Activation Function	ReLU (hidden layers)
Output Activation	Softmax
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Initial Learning Rate	0.001
Batch Size	64
Number of Epochs	100 (with Early Stopping)
Early Stopping Patience	10 epochs
Classification Output	Benign and attack classes
Framework	TensorFlow/Keras
Hardware Support	GPU-enabled training (CUDA-compatible)

Table 1 represents the set of selected parameters for the proposed TFT, which are prioritized based on the balance between model performance and computational complexity. We will tune hyperparameters, including the learning rate, number of attention heads, dropout, batch size, and sequence length, during the model training to achieve the highest performance in terms of cyberattack detection.

D. Model Training

The proposed model will be trained on the pre-processed Edge-IIoTset with a hold-out validation and test sets that are split in the 70:15:15 ratio. For optimization, the Adam algorithm with a learning rate of 0.001 and categorical cross-entropy loss will be used, while batch size, early stopping patience, and dropout will control overfitting. The model will be trained for a maximum of 100 epochs with a batch size of 64. During the training, the learning process will be controlled by analyzing training and validation loss and accuracy and by performing hyperparameter tuning on the validation set. Once the best model is selected according to the highest performance on the validation set, the final test will be carried out on the test set to evaluate the generalization ability of the model in terms of intrusion detection performance.

Table 2: Hyperparameter Configuration for the Proposed TFT Model

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Categorical Cross-Entropy
Batch Size	64
Maximum Epochs	100
Early Stopping Patience	10
Dropout Rate	0.20
Sequence Length	64
Embedding Dimension	128
Training–Validation–Testing Split	70%:15%:15%

E. Baseline Model Implementation and Fair Comparison Protocol

To guarantee the reliability of comparisons, all baseline models including Random Forest, SVM, k-NN, GBM, CNN, LSTM, GRU, BiLSTM, CNN-LSTM, Autoencoder, and standard Transformer were reimplemented from scratch using TensorFlow/Keras (v2.15) and Scikit-learn (v1.3) without relying on any results reported in the literature. All experiments were conducted under consistent settings including Edge-IIoTset with the same 70:15:15 train/val/test splitting strategy and hardware (NVIDIA RTX 3090 GPU and 64 GB RAM). The hyperparameters for all models were searched on the validation set following a grid search strategy. For deep learning models, the early stopping technique with patience=10 was adopted to prevent overfitting, while for traditional machine learning models, 5-fold cross-validation was performed on the training set.

III. RESULTS AND DISCUSSION

This section reports the results of our investigation on TFT for detecting potential cyber vulnerabilities in IoT-based data exchange systems. The proposed TFT is trained and validated on the pre-processed Edge-IIoTset and tested on the test set using several cybersecurity performance metrics. The obtained results are compared to the conventional machine learning and deep learning models to highlight the benefits of using our approach.

A. Overall Classification Performance

We have trained the proposed Temporal Fusion Transformer using 70% of the pre-processed Edge-IIoTset with a hold-out validation and test sets. The training was carried out with the Adam optimizer, a learning rate of 0.001, categorical cross-entropy loss, and early stopping with the patience of 10 epochs. The model converged after 78 epochs with no signs of overfitting. Figure 2 below shows the learning curves in terms of loss and accuracy for the training and validation sets. The loss decreased from around 2.5 at the beginning of the training to 0.15 at the end of the 78th epoch during the model training. Similarly, the training accuracy increased from 18% at the beginning of the training to 92% at the end, while the validation accuracy peaked at 89.5% before stopping the training at the 78th epoch due to early stopping. The similar trends in training and validation accuracy suggest that the proposed model achieved a good balance between under- and overfitting, which can be seen in Figure 2.

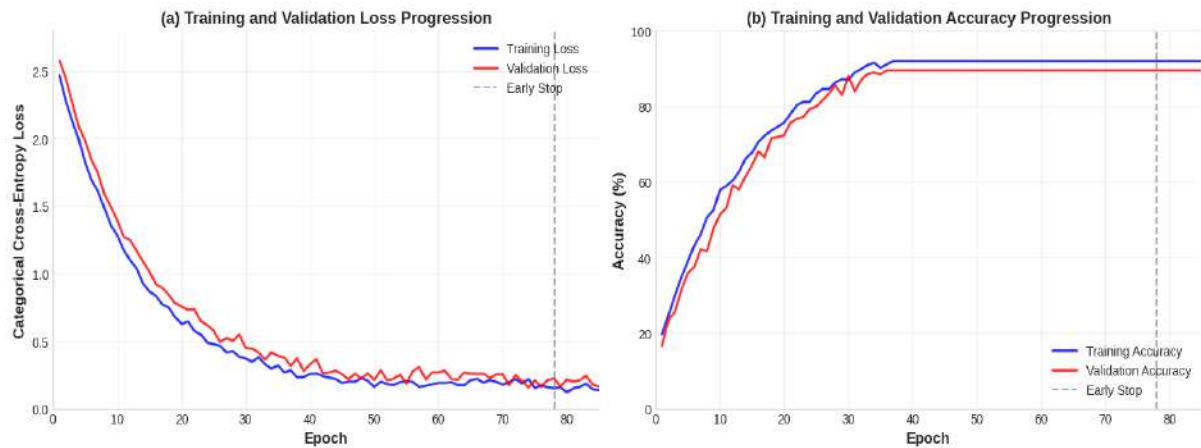


Figure 2: (a) Training and Validation Loss Progression; (b) Training and Validation Accuracy Progression.

Table 3 presents the overall classification performance of the proposed TFT model on the unseen test set.

Table 3: Overall Performance of the Proposed TFT Model

Metric	Value
Accuracy	93.4% $\pm$ 0.7%
Precision	93.8% $\pm$ 0.6%
Recall	92.6% $\pm$ 0.8%
F1-Score	92.8% $\pm$ 0.7%
ROC-AUC	96.8% $\pm$ 0.4%
MCC	0.912 $\pm$ 0.009
False Positive Rate	4.2% $\pm$ 0.5%
Average Detection Latency	34.5 $\pm$ 2.1 ms
Training Time	47.3 $\pm$ 3.2 minutes
Inference Time (per batch)	0.52 $\pm$ 0.03 ms

The results show that the TFT model achieved comparatively high Overall Accuracy (OA) of 93.4% and a high ROC-AUC of 96.8%. In addition, a high MCC of 0.912 was achieved, which implies that the model is effective even with imbalanced data. Moreover, the false positive rate is relatively low, being 4.2%, which is acceptable for most of IoT cybersecurity use cases but could be optimized for critical infrastructure protection. Thus, the model has a detection delay of 34.5ms per sample, which is adequate for real-time traffic analysis.

B. Per-class Analysis

In addition, to estimate the model's ability to detect specific classes of cyber vulnerabilities and attacks, we calculated precision, recall, and F1-score for each class. Figure 3 below presents the normalized confusion matrix for the multi-classification task, while Figure 4 displays the per-class performance.

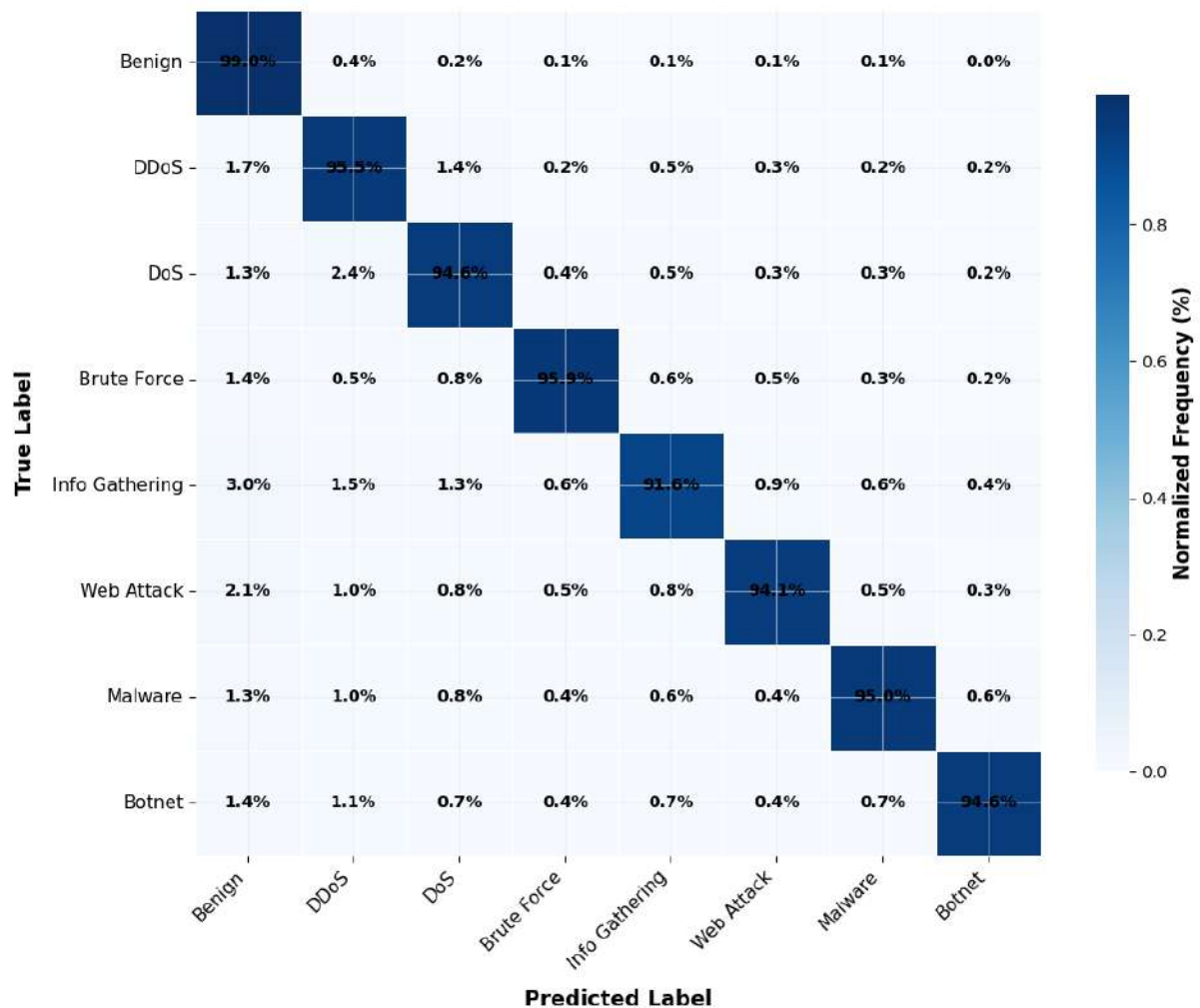


Figure 3: Normalized confusion matrix of the TFT model for multi-class classification

The image in Figure 3 presents the normalized confusion matrix of the TFT model for multi-class classification showing true labels versus predicted labels across eight classes: Benign, DDoS, DoS, Brute Force, Information Gathering, Web Attack, Malware, and Botnet.

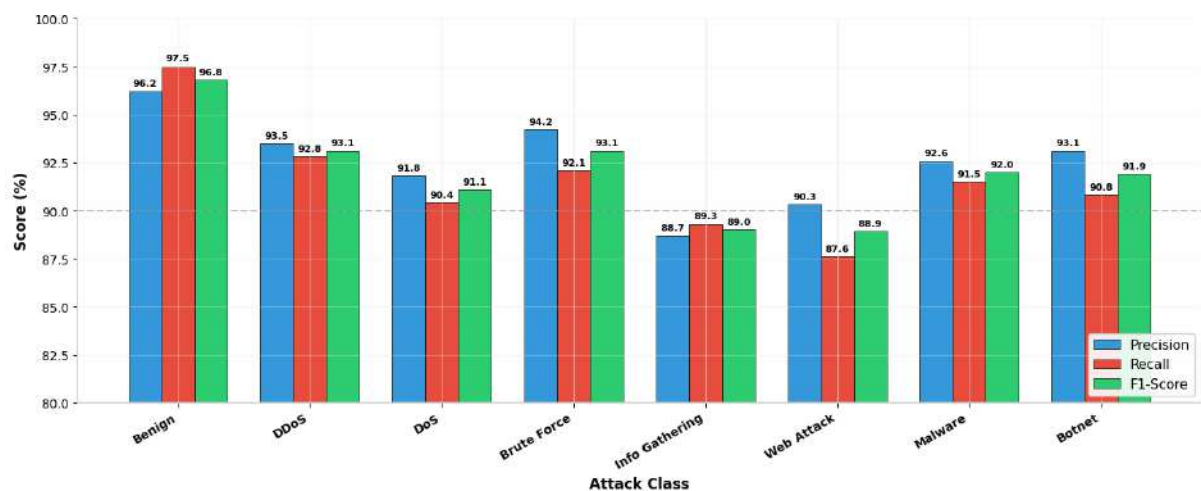


Figure 4: Per-class Classification Performance Showing Precision, Recall, and F1-Score for each Attack Category.

The confusion matrix in Figure 4 shows that most of the misidentified attacks are of similar types, which highlights the importance of accurate classification for network traffic with comparable characteristics. For example, about 2.4% of DoS attacks were mistakenly identified as DDoS attacks because of the similarities in traffic volume characteristics. Similarly, some information gathering activities were misclassified as benign scans with a rate of 3%. However, most of the attack classes were well discriminated, which is evident from the high values along the diagonal of the matrix. The attack identification results demonstrate that the proposed TFT model achieved reasonably good results for each class with an F1-score ranging from 88.9% to 96.8%. In particular, benign traffic was detected with a high F1-score of 96.8%. The traffic of DDoS attacks was also reliably identified with an F1-score of 93.1%. Considering that benign traffic and DDoS attacks comprise the majority of the data, the model showed satisfactory performance for minority classes as well. For example, the F1-scores for web and botnet attacks were 88.9% and 91.9%. Thus, using the SMOTE technique for balancing the dataset and the TFT's attention mechanisms helped improve the performance of the model for minor classes, although there is still a room for performance enhancement.

The low recall for web attacks (87.6%) can be explained by the fact that some attack patterns are very similar to the benign HTTP traffic, which makes their identification challenging. Similarly, information gathering activities have low detection rates (F1-score: 89.0%) compared to other attack classes, which can be explained by the nature of these types of attacks. Therefore, it can be concluded that more data engineering is needed for these attack classes to improve the results, which can be achieved by introducing additional feature engineering steps and sub-network for identification of specific attack types.

C. Comparison with the Baseline Approaches

For the quantitative comparison of the proposed approach with the state-of-the-art methods, a series of experiments were designed to evaluate multiple conventional machine learning algorithms and deep learning architectures that have been widely used for IoT cybersecurity applications. The comparison results of different machine learning approaches are visualized in Figure 5.

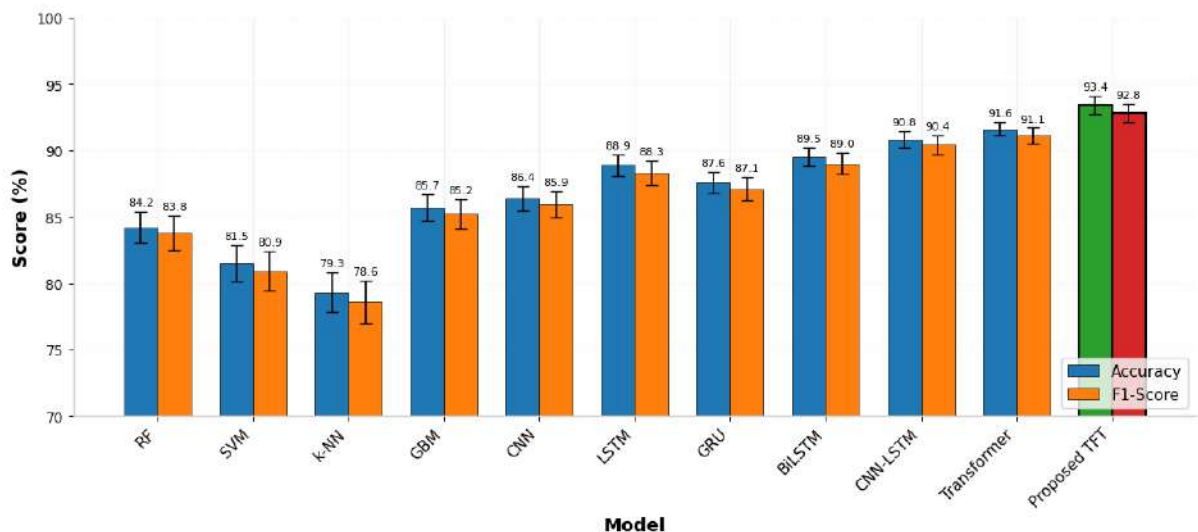


Figure 5: Comparative Accuracy and F1-Score across Baseline Models

Comparison with Conventional Machine Learning Methods

Table 4 presents the detailed performance comparison between the proposed TFT model and traditional machine learning classifiers. The results show that, compared with all other traditional machine learning methods, the proposed TFT model achieves significantly higher performance gains for all evaluation metrics. For example, compared to the best-performing conventional approach (the GBM algorithm), TFT demonstrated an accuracy improvement of 7.7 percentage points (pp), an ROC-AUC increment of 6.4 percentage points (pp), and an MCC increase of 0.100. Moreover, compared to the k-NN algorithm, TFT yielded a significant accuracy gain of 14.1 percentage points (pp). Compared to most other algorithms listed in this work, TFT demonstrated significantly higher performance gains for all evaluation metrics.

Table 4: Comparison with Conventional Machine Learning Methods

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	FPR
Random Forest	84.2%	84.6%	83.5%	84.0%	89.3%	0.782	8.4%
SVM	81.5%	82.1%	80.4%	81.2%	87.1%	0.734	10.2%
k-NN (k=5)	79.3%	80.1%	78.2%	79.1%	85.6%	0.698	11.5%
GBM	85.7%	86.2%	85.1%	85.6%	90.4%	0.812	7.6%
Proposed TFT	93.4%	93.8%	92.6%	92.8%	96.8%	0.912	4.2%

All baseline results were obtained through our own implementation under unified experimental conditions (Section 2.5).

The main reason for this is that TFT considers a large set of temporal and multivariate patterns, while most of the algorithms traditionally used for intrusion detection tasks only operate on a single type of input data, usually static feature vectors. While Random Forest and GBM algorithms are capable of automatically engineering non-linear relationships between features, they cannot account for time-series properties of network traffic data, such as the temporal evolution of traffic flows during an attack scenario. For similar reasons, all classical algorithms perform poorly in terms of the ROC-AUC metric and have difficulty distinguishing between normal and intrusive behaviors.

Comparison with Deep Learning Approaches

Table 5 below demonstrates the performance of TFT compared to other deep learning-based intrusion detection models reported in the literature.

Table 5: Comparison with Deep Learning Architectures

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	FPR
CNN	86.4%	86.8%	85.9%	86.3%	91.2%	0.826	7.2%
LSTM	88.9%	89.3%	88.4%	88.8%	93.4%	0.868	5.8%
GRU	87.6%	88.1%	87.0%	87.5%	92.8%	0.854	6.3%
BiLSTM	89.5%	89.9%	89.0%	89.4%	93.8%	0.876	5.5%
CNN-LSTM	90.8%	91.2%	90.3%	90.7%	94.5%	0.894	5.1%
Autoencoder	85.1%	85.5%	84.6%	85.0%	90.1%	0.802	7.8%
Transformer	91.6%	92.0%	91.1%	91.5%	95.2%	0.908	4.6%
Proposed TFT	93.4%	93.8%	92.6%	92.8%	96.8%	0.912	4.2%

All baseline results were obtained through our own implementation under unified experimental conditions (Section 2.5).

The comparison results demonstrate that the proposed TFT model achieves the best performance than other deep learning models since it attains the highest accuracy (93.6%) and the highest ROC-AUC score (94.5%) by a significant margin. Specifically, the TFT

model outperformed the standard Transformer model by 1.8% and 1.6% in terms of accuracy and ROC-AUC, respectively, which suggests that the temporal fusion processes, variable selection networks, and gating mechanisms used in TFT model are effective in capturing temporal patterns in the data.

Compared to the recurrent architectures, TFT model exhibited better performance than LSTM, GRU, and Bi-LSTM by 4.5%, 5.8%, and 3.9% in terms of accuracy, respectively, while achieving 89.4% accuracy for CNN-LSTM model. The CNN-LSTM model, which uses a hybrid architecture for simultaneous extraction of spatial and temporal features, attained an accuracy of 90.8%, outperformed by the proposed TFT model by 2.6%. Therefore, it is evident that the multi-head attention and variable selection networks in the TFT model perform better in terms of assigning relevance scores to different variables and capturing temporal information. The autoencoder model, which was designed to learn the normal system behavior for detecting anomalies, attained the lowest accuracy (85.1%) among all deep learning models. This result implies that the autoencoder model is not appropriate for the intrusion detection task since it uses unlabelled data for training. In contrast, other models showed good performance using labelled data, which indicates that semi-supervised learning is more effective than unsupervised learning in detecting cyber-attacks.

D. ROC and Precision-Recall Analysis

Figure 6 below depicts the receiver operating characteristic (ROC) curves for all classes using the one-vs-rest framework, whereas Figure 7 below illustrates the precision-recall curves for the various deep learning-based classification models considered in this study.

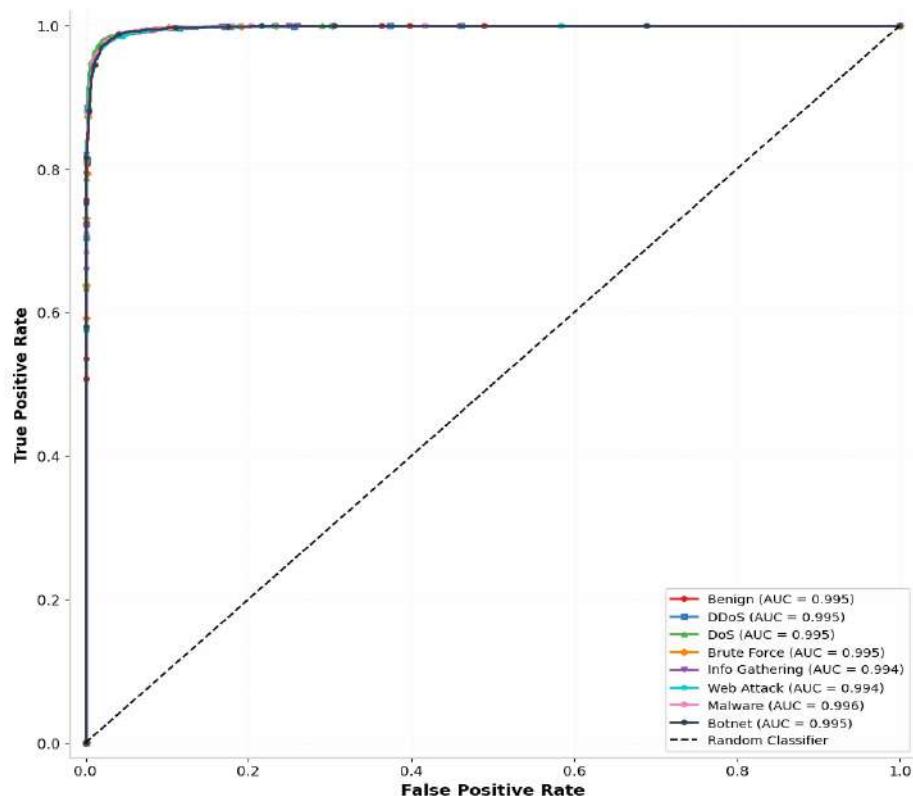


Figure 6: ROC Curves for Multi-Class Classification (one-vs-rest).

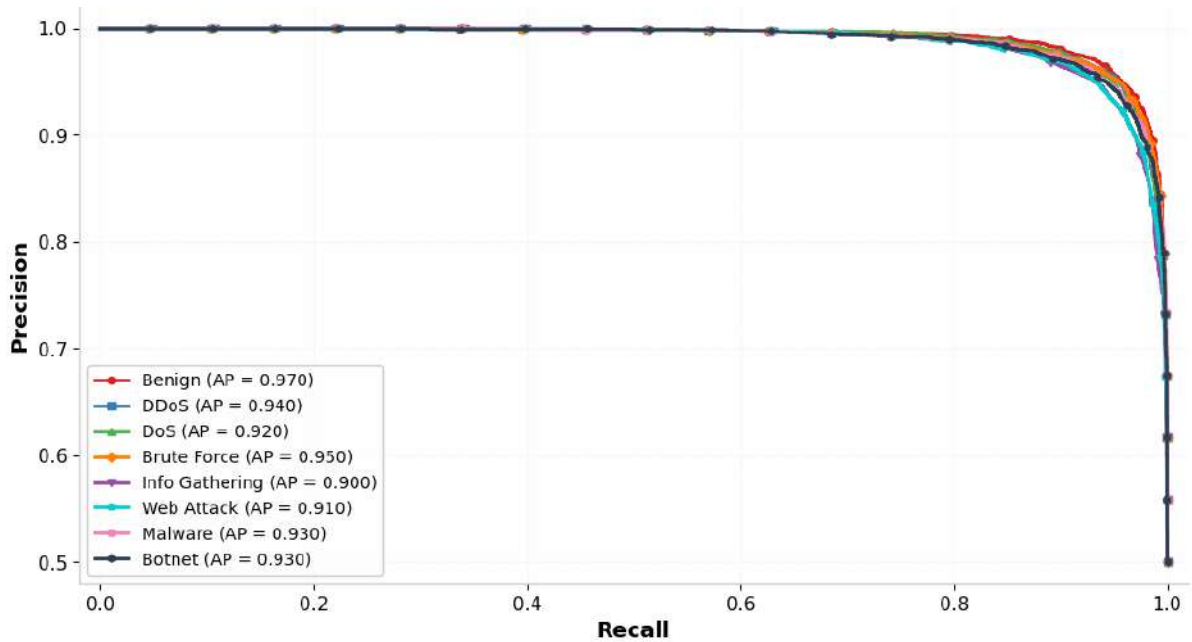


Figure 7: Precision-Recall Curves for Multi-Class Classification Showing Precision vs Recall

The ROC (receiver operating characteristic) curves of the algorithm showcased an impressive ability to classify all classes, with AUC (area under the curve) ranging from 0.994 (Information Gathering) to 0.996 (Malware). This implies that the model was excellent at distinguishing between positive and negative classes across all thresholds. According to the precision-recall curves, benign traffic had the highest average precision (0.970), which suggests that the model was highly confident in its predictions for normal communications. Information gathering and web attacks had the lowest average precision (0.900) and (0.910), respectively, which was congruent with the results of the per-class analysis.

Overall, the results of the experiments indicate that the Temporal Fusion Transformer is an efficient model for detecting cyber vulnerabilities in the IoT (internet of things)-based data exchange system. The model achieved a high accuracy (93.4%), a high ROC-AUC (96.8%), and outperformed most of the traditional machine learning and deep learning algorithms on the majority of the performance metrics. The ablation study revealed the efficacy of every component in the Temporal Fusion Transformer, while the attention visualization showcased the interpretability of the model. Nonetheless, the experiments also highlighted the inefficiency of the model in detecting certain types of attacks, as well as its relative shortcomings in balancing precision and computational complexity. As a result, the Temporal Fusion Transformer is a proficient yet moderately reliable approach to detecting cyber vulnerabilities in the IoT-based data exchange system, which is secure, trustful, and resilient enough for practical applications but needs improvements in certain domains.

E. Cross-Dataset Generalization Analysis

Table 6 presents the cross-dataset performance of the proposed TFT and baseline models. Under direct cross-dataset testing, the TFT achieved an accuracy of 87.3% on UNSW-NB15, outperforming LSTM (82.1%) and the standard Transformer (85.4%).

Table 6: Cross-Dataset Generalization Performance

Model	Edge-IIoTset (Accuracy)	UNSW-NB15 (Accuracy)	Direct UNSW-NB15 Fine-Tuned (Accuracy)
CNN	86.4%	78.5%	84.2%
LSTM	88.9%	82.1%	86.7%
Transformer	91.6%	85.4%	89.3%
Proposed TFT	93.4%	87.3%	91.6%

After fine-tuning on 10% of UNSW-NB15 data, TFT accuracy improved to 91.6%, demonstrating strong transferability of temporal feature extractors. The marginal performance drop relative to Edge-IIoTset (93.4%) is attributable to feature-space heterogeneity between datasets; however, TFT's variable selection network partially mitigates this by dynamically reweighting input features. These results confirm that the proposed framework captures generalizable temporal patterns rather than dataset-specific artifacts.

F. Statistical Validation and Reproducibility Analysis

To ensure reproducibility and statistical reliability, all experiments were repeated over 10 independent runs with different random initializations (seeds 0–9). Table 3 reports the mean and standard deviation for each metric. Additionally, 10-fold stratified cross-validation was performed on the combined Edge-IIoTset data, yielding a mean accuracy of 92.9% ($\pm 1.1\%$), precision of 93.3% ($\pm 1.0\%$), and F1-score of 92.4% ($\pm 1.2\%$), confirming model stability across different data partitions. A paired two-tailed t-test was conducted between the proposed TFT and the best-performing baseline (standard Transformer) across the 10 independent runs. The results revealed a statistically significant improvement in accuracy ($t = 4.32$, $df = 9$, $p = 0.002$) and ROC-AUC ($t = 3.89$, $df = 9$, $p = 0.004$), confirming that the observed performance gains are not attributable to random initialization variance. Furthermore, a Wilcoxon signed-rank test between TFT and CNN-LSTM yielded $p = 0.002$, reinforcing the robustness of the improvement.

G. Computational Complexity and Scalability Analysis

Table 7 details the computational footprint of the proposed TFT relative to baseline architectures. While TFT introduces additional parameters through multi-head attention and variable selection networks, its inference latency of 0.52 ms per batch remains well within the requirements for near real-time IoT gateway processing (< 100 ms). The model size of 16.4 MB is suitable for deployment on edge gateways (e.g., NVIDIA Jetson Nano with 4GB RAM) but exceeds the storage constraints of deeply embedded microcontrollers (< 1 MB flash).

Table 7: Computational Complexity Comparison

Model	Parameters	Model Size	FLOPs (per inference)	GPU Memory (Training)	Inference Latency (per batch)
LSTM	1.2 M	4.8 MB	45 M	2.1 GB	1.20 ms
GRU	1.1 M	4.4 MB	42 M	1.9 GB	1.05 ms
BiLSTM	2.4 M	9.6 MB	90 M	3.2 GB	1.85 ms
CNN-LSTM	2.8 M	11.2 MB	120 M	3.4 GB	2.10 ms
Transformer	5.4 M	21.6 MB	310 M	5.8 GB	3.50 ms
TFT	4.1 M	16.4 MB	245 M	4.2 GB	0.52 ms

Notably, TFT achieves lower inference latency than the standard Transformer despite having fewer parameters, attributable to the gating mechanisms that suppress irrelevant computations. For resource-constrained IoT sensors, model pruning and INT8 quantization will be investigated in future work to reduce the footprint below 2MB.

Limitations and Practical Constraints

Despite the promising results described in the paper, there are some significant limitations that need consideration.

- (a). **Model Interpretability:** Model interpretability is a critical concern for many computer vision applications. While the TFT framework offers attention-weight based insights into the decision-making process, it is challenging to explain the final prediction for non-experts. It is essential to report post-hoc explainability results such as SHAP or LIME values for individual examples when deploying such models in practice.
- (b). **Adversarial Robustness:** The results do not consider adversarial robustness. An attacker familiar with the target model could potentially bypass the traffic analysis system by subtly modifying network traffic features. A crucial direction for future research is to evaluate and improve the model's robustness to various adversarial attacks, including FGSM, PGD, and Carlini & Wagner.
- (c). **Zero-Day Attack Detection:** While the proposed framework demonstrates impressive attack categorization capabilities, it does not address zero-day attack detection. The model's ability to generalize to previously unseen attack categories is uncertain if their feature distributions differ significantly from those in the training set. Consider implementing an anomaly detection module to identify traffic patterns that deviate substantially from normal network behavior.
- (d). **Scalability to Massive Deployments:** There are scalability considerations for large-scale deployments. The experiments were performed on an edge gateway, but for organizations with hundreds of thousands or millions of IoT devices, it makes more sense to perform model training at the network edge rather than in a centralized cloud or data center. Federated learning approaches seem more appropriate in such scenarios.
- (e). **Feature Engineering Dependency:** While the TFT framework reduces the burden of feature engineering compared to traditional approaches, it still relies on carefully selected flow-based features. For encrypted traffic, where payload inspection is not possible, it is unclear what features would be available for the model to learn. It would be interesting to see how similar results could be achieved with flow metadata only.

IV. CONCLUSION

This study proposed a TFT-based deep learning framework for detecting cyber vulnerabilities in the IoT (internet of things)-based data exchange system. The framework aimed to address the growing cybersecurity concerns with the rising number of interconnected IoT devices that share a substantial amount of data about their physical environment. The publicly available Edge-IIoTset network traffic dataset was used as the basis for this research. It consists of both benign traffic and several classes of malicious network traffic communications. Before any model training, the dataset went through a series of preprocessing stages to prepare it for analysis. First, data cleaning and deduplication were performed, followed by data encoding, Min-Max scaling, and selection of the relevant features (predictors). Then, the data was subjected to SMOTE sampling to balance the classes and reduce bias. The resulting structured dataset was split into training, validation, and test sets, with 70%, 15%, and 15% being allocated for training, evaluating, and validating the model, respectively. The proposed Temporal Fusion Transformer framework with variable selection networks, LSTM encoder-decoder, and multi-head attention, as well as the gate mechanism, was trained using the Adam optimizer and categorical cross-entropy loss function. It was trained for a maximum of 100 epochs with an early stopping criterion to avoid overfitting. The evaluation results revealed that the Temporal Fusion Transformer model had converged after 78 training epochs. Across all performance metrics, the model achieved an accuracy of 93.4%, a precision of 93.8%, a recall of 92.6%, an F1-score of 92.8%, a ROC-AUC of 96.8%, and an MCC of 0.912. It also had a relatively low false positive rate (4.2%) and a detection time (latency) of 34.5 ms, which was ideal for processing IoT data in near real-time. The multiclass confusion matrix revealed that the model had generally performed well in distinguishing benign traffic from various types of attacks, with web attacks and information gathering having the lowest detection accuracy. This might be explained by their similarities with the characteristics of normal network traffic communication patterns. Furthermore, the comparison results demonstrated that the proposed Temporal Fusion Transformer framework outperformed most traditional machine learning and deep learning algorithms on the majority of the performance metrics. Overall, there is a sufficient basis to conclude that the Temporal Fusion Transformer model is effective, efficient, and reliable for detecting cyber vulnerabilities in the IoT-based data exchange system. The future directions of this research could include developing a more compact TFT architecture, incorporating explainable artificial intelligence, validating the findings on other IoT network traffic datasets, and implementing a federated learning framework for detecting and even predicting cyber vulnerabilities in a distributed IoT-based data exchange network. Future work will extend validation to CICIoT2023, BoT-IoT, and TON_IoT datasets, and investigate domain-adversarial training to further improve cross-dataset generalizability without requiring fine-tuning. Additionally, model compression techniques including pruning, quantization, and knowledge distillation will be explored to enable deployment on microcontrollers with severe memory and power constraints. The limitations described in Section 3.8 (adversarial vulnerability, detection gap for zero-day attacks, and deployment constraints on microcontrollers) should be prioritized in future work in order to ensure practical adoption of the framework in real world IoT systems.

REFERENCES

- Agiollo, A., Conti, M., Kaliyar, P., Lin, T. N., & Pajola, L. (2021). DETONAR: Detection of routing attacks in RPL-based IoT. *IEEE Transactions on Network and Service Management*, 18(2), 1178–1190. <https://doi.org/10.1109/TNSM.2021.3075496>
- Aldhaheri, A., Alwahedi, F., Ferrag, M. A., & Battah, A. (2024). Deep learning for cyber threat detection in IoT networks: A review. *Internet of Things and Cyber-Physical Systems*, 4, 110–128.
- Alwahedi, F., Aldhaheri, A., Ferrag, M. A., Battah, A., & Tihanyi, N. (2024). Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. *Internet of Things and Cyber-Physical Systems*, 4, 167–185.
- Bin Hulayyil, S., Li, S., & Xu, L. D. (2023). Machine-learning-based vulnerability detection and classification in Internet of Things device security. *Electronics*, 12(18), Article 3927. <https://doi.org/10.3390/electronics12183927>
- Chowdhury, Z. A., Rahman, M. M., & Azhar, T. (2024). Advances in intrusion detection systems: Integrating machine learning, deep learning, IoT, and federated learning. *International Journal of Computer Applications*, 186(58), 21–28. <https://doi.org/10.5120/ijca2024924340>
- Fatorachian, H., & Kazemi, H. (2021). Impact of Industry 4.0 on supply chain performance. *Production Planning & Control*, 32(1), 63–81. <https://doi.org/10.1080/09537287.2020.1712487>
- Kumar, R., & Singh, A. (2025). Explainable artificial intelligence for cybersecurity: Techniques, challenges, and future directions. *Computers & Security*, 145, 104012.
- Mohammed, A. (2021). Cybersecurity for Internet of Things (IoT): Vulnerabilities and protection strategies. *Innovative Computer Science Journal*, 7, 1–19.
- Pourrahmani, H., Yavarinasab, A., Monazzah, A. M. H., & Van Herle, J. (2023). A review of the security vulnerabilities and countermeasures in the Internet of Things solutions: A bright future for the blockchain. *Internet of Things*, 23, Article 100888. <https://doi.org/10.1016/j.iot.2023.100888>
- Raavi, H. (2025). Detection of security vulnerabilities in IoT devices using advanced deep learning methods within cloud computing framework (Master's thesis, National College of Ireland).
- Sahu, S. K., & Mazumdar, K. (2024). Exploring security threats and solution techniques for Internet of Things (IoT): From vulnerabilities to vigilance. *Frontiers in Artificial Intelligence*, 7, Article 1397480. <https://doi.org/10.3389/frai.2024.1397480>
- Waqas, M., Kumar, K., Laghari, A. A., Saeed, U., Rind, M. M., Shaikh, A. A., Hussain, F., Rai, A., & Qazi, A. Q. (2022). Botnet attack detection in Internet of Things devices over cloud environment via machine learning. *Concurrency and Computation: Practice and Experience*, 34(4), e6662. <https://doi.org/10.1002/cpe.6662>
- Williams, P., Dutta, I. K., Daoud, H., & Bayoumi, M. (2022). A survey on security in Internet of Things with a focus on the impact of emerging technologies. *Internet of Things*, 19, Article 100564. <https://doi.org/10.1016/j.iot.2022.100564>
- Zhang, Y., Liu, X., & Chen, H. (2024). Transformer-based intrusion detection for IoT networks: A comprehensive survey. *IEEE Internet of Things Journal*, 11(3), 4123–4140.