ISSN 2723-6471

Incorporate Transformer-Based Models for Anomaly Detection

Deshinta Arrova Dewi[1], Harprith Kaur Rajinder Singh[2], Jeyarani Periasamy[3], Tri Basuki Kurniawan[4,*], Henderi[5], M. Said Hasibuan[6], Yogeswaran Nathan[7]

[1,2,3]*Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

[4]*Faculty of Science and Technology, Universitas Bina Darma, Palembang, Indonesia*

[5]*Department of Informatics Engineering, University of Raharja, Indonesia*

[6]*Faculty Computer Science, Institute Informatics and Business Darmajaya, Bandar Lampung, Indonesia*

[7]*School of Computing, Asia Pacific University, Malaysia*

(Received: November 28, 2024; Revised: January 5, 2025; Accepted: April 10, 2025; Available online: July 13, 2025)

Abstract

This paper explores the effectiveness of Transformer-based models, specifically the Time-Series Transformer (TST) and Temporal Fusion Transformer (TFT), for anomaly detection in streaming data. We review related work on anomaly detection models, highlighting traditional methods' limitations in speed, accuracy, and scalability. While LSTM Autoencoders are known for their ability to capture temporal patterns, they suffer from high memory consumption and slower inference times. Though efficient in terms of memory usage, the Matrix Profile provides lower performance in detecting anomalies. To address these challenges, we propose using Transformer-based models, which leverage the self-attention mechanism to capture long-range dependencies in data, process sequences in parallel, and achieve superior performance in both accuracy and efficiency. Our experiments show that TFT outperforms the other models with an F1-score of 0.92 and a Precision-Recall AUC of 0.71, demonstrating significant improvements in anomaly detection. The TST model also shows competitive performance with an F1-score of 0.88 and Precision-Recall AUC of 0.68, offering a more efficient alternative to LSTMs. The results underscore that Transformer models, particularly TST and TFT, provide a robust solution for anomaly detection in real-time applications, offering improved performance, faster inference times, and lower memory usage than traditional models. In conclusion, Transformer-based models stand out as the most effective and scalable solution for large-scale, real-time anomaly detection in streaming time-series data, paving the way for their broader application across various industries. Future work will further focus on optimizing these models and exploring hybrid approaches to enhance detection capabilities and real-time performance.

Keywords: Anomaly Detection, Time-Series Forecasting, Transformer Models, Temporal Fusion Transformer (TFT), Streaming Data, Process Innovation

1. Introduction

The rapid increase in real-time data streams across various domains, such as finance, healthcare, and transportation, has urgently needed accurate and scalable anomaly detection methods. Traditional approaches, including statistical models and machine learning algorithms, often struggle with complex dependencies, high dimensionality, and the evolving nature of streaming data [1]. While Long Short-Term Memory (LSTM) networks have successfully captured sequential dependencies, they face limitations in handling long-range correlations efficiently due to their recurrent nature and vanishing gradient problem [2].

Recent advancements in deep learning have introduced Transformer-based models, such as TST [3] and TFT [4], which offer significant improvements over LSTMs for sequence modelling. These models leverage self-attention mechanisms to capture long-term dependencies while efficiently processing large-scale streaming data in parallel [5]. This study explores the application of Transformer-based models for real-time anomaly detection in streaming environments, using the NYC Taxi dataset, to demonstrate their effectiveness compared to LSTM autoencoders. We propose utilizing TST and TFT for anomaly detection in streaming data. These models are designed to address the limitations of LSTMs

***Corresponding author: Tri Basuki Kurniawan (tribasukikurniawan@binadarma.ac.id)**
DOI: https://doi.org/10.47738/jads.v6i3.762

by capturing long-range dependencies using self-attention mechanisms, processing sequences in parallel for improved efficiency, and dynamically adapting to changes in data patterns to enhance anomaly detection performance. The TST model will process raw time-series data, leveraging self-attention to identify anomalies without predefined feature engineering [6]. The TFT model will integrate additional context features, such as temporal trends and categorical information, for enhanced anomaly detection [7].

The NYC Taxi dataset, containing taxi trip records over six months, will be used for model evaluation. This dataset is widely used for anomaly detection research due to its inherent time-series patterns and anomalies related to holidays, weather events, and transportation disruptions [8]. To benchmark performance, baseline models will include the LSTM Autoencoder, previously used in the existing study for anomaly detection, and the Matrix Profile Algorithm, serving as a statistical baseline [9]. The proposed models consist of the TST, which utilizes multi-head self-attention for detecting anomalous sequences, and the TFT, which integrates multiple features and dynamic relationships for anomaly detection.

The evaluation process will focus on several key metrics. The F1-score will measure precision and recall trade-offs, while Precision-Recall AUC will evaluate anomaly detection effectiveness [10]. Inference time will be assessed to determine real-time processing efficiency, and memory usage will be analysed to compare the computational overhead of each model [11]. Implementation details involve using PyTorch with Hugging Face's Transformer library, training models with the Adam optimiser, and fine-tuning learning rates using Optuna [12]. Anomalies will be detected using reconstruction errors for autoencoders and attention scores for Transformers. To benchmark performance, the models will be deployed on an NVIDIA GeForce RTX 3080 GPU.

This study explores the advantages of Transformer-based models over traditional LSTMs for real-time anomaly detection in streaming data. By leveraging self-attention mechanisms, these models improve accuracy, scalability, and efficiency [12]. The experimental results will highlight the strengths of TST and TFT in detecting anomalies, paving the way for future research in deploying Transformers for real-time applications across various industries.

2. Literature Review

Anomaly detection in time-series data has traditionally relied on statistical models such as ARIMA, Holt-Winters, and Gaussian processes, alongside classical machine learning algorithms like k-nearest neighbors, support vector machines, and isolation forests. While effective in stationary or low-dimensional settings, these methods are often inadequate for high-dimensional, non-linear, and dynamic environments due to their limited capacity to model complex temporal dependencies and adapt to evolving data distributions.

The emergence of deep learning brought significant progress, particularly through Long Short-Term Memory (LSTM) networks. LSTM architectures are capable of learning long-term dependencies in sequential data due to their internal gating mechanisms—input, forget, and output gates—that selectively update and retain relevant information across time steps. This structure allows LSTMs to avoid vanishing gradient issues common in vanilla RNNs and to effectively model long-range temporal relationships. The core update equations are typically formulated as:

$$Forget\ gate: f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$\text{Input gate: } i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\text{Candidate memory: } \tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$\text{Output gate: } o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

$$\text{Hidden state: } h_t = o_t \cdot \tanh(C_t) \quad (4)$$

Here, xt represents the input at time t, ht the hidden state, and Ct the memory cell. The sigmoid function σ regulates gate activation, while the tanh function introduces non-linearity and bounds the memory values.

Despite their success, LSTMs exhibit several limitations, particularly for real-time and large-scale applications. Their sequential nature prevents full parallelization during training, leading to slower convergence times. Additionally, their

memory requirements scale poorly with long sequences, limiting their feasibility in environments with strict latency and resource constraints [2], [12].

To overcome these challenges, Transformer-based architectures have gained attention due to their superior scalability and modeling flexibility. Unlike RNNs, Transformers dispense with recurrence entirely, instead using multi-head self-attention mechanisms to model relationships between all-time steps in parallel. This enables them to efficiently capture both short-term and long-range dependencies across time without being constrained by sequence order [3], [5].

Recent advancements in Transformer-based models, such as the TST and TFT, have achieved state-of-the-art results in forecasting and anomaly detection across domains like finance, healthcare, and industrial monitoring [4], [6], [7]. These models offer improved scalability and interpretability, often incorporating additional components such as gating mechanisms, variable selection networks, and static covariate encoders to enhance prediction fidelity and domain adaptation [7].

Furthermore, comparative studies show that Transformer-based models outperform LSTM-based Autoencoders in anomaly detection tasks, achieving higher F1-scores, precision-recall AUC, and inference efficiency. These advantages are particularly notable in noisy, high-dimensional, and multi-sensor data streams [9], [10]. The attention mechanism's ability to focus on relevant portions of the sequence allows it to detect subtle or irregular deviations that traditional methods might overlook.

The increasing adoption of Transformer architectures marks a paradigm shift in sequential modeling. Rather than replacing LSTM models entirely, hybrid approaches that integrate attention-based mechanisms with recurrent or statistical components are emerging. These combinations aim to leverage the contextual awareness of LSTMs with the global dependency modeling and parallelization benefits of Transformers, creating more adaptive and robust anomaly detection pipelines [11]. As research in this domain progresses, future work is likely to focus on improving the efficiency, robustness, and explainability of Transformer-based systems, particularly for deployment in streaming, imbalanced, or real-time environments.

3. Methodology

We propose utilizing TST and TFT for anomaly detection in streaming data. These models are designed to address the limitations of LSTMs by capturing long-range dependencies using self-attention mechanisms, processing sequences in parallel for improved efficiency, and dynamically adapting to changes in data patterns to enhance anomaly detection performance. The TST model will process raw time-series data, leveraging self-attention to identify anomalies without predefined feature engineering [13]. The TFT model will integrate additional context features, such as temporal trends and categorical information, for enhanced anomaly detection [14]. The methodology of this study consists of four major components: data preprocessing, model implementation, training process, and evaluation metrics. Figure 1 illustrates the complete research workflow for implementing Transformer-based models, specifically the TST and TFT, for real-time anomaly detection in streaming data. The process begins by defining the research objective and proposing TST and TFT as solutions to overcome the limitations of traditional models such as LSTMs. The workflow then proceeds to data preprocessing using the NYC Taxi dataset, where missing values are handled through interpolation, features are normalized using Min-Max scaling, and temporal context features like hour and holiday indicators are added before splitting the dataset into training and testing subsets. Next, three modeling paths are explored in parallel: TST, TFT, and two baselines, namely LSTM Autoencoders and the Matrix Profile algorithm. These models are trained using the PyTorch framework and optimized with the Adam optimizer and Optuna-based hyperparameter tuning, with early stopping applied to prevent overfitting. After training, model performance is evaluated using F1 score, precision-recall AUC, inference time, and memory usage. Based on the evaluation results, the best-performing model is either deployed directly or, if performance is unsatisfactory, the workflow loops back for model refinement or hybrid design. The final stage focuses on deployment, where the selected model is optimized using quantization techniques and integrated into a real-time detection system, resulting in a robust and efficient solution for anomaly detection in high-throughput, time-sensitive environments.

 ISSN 2723-6471

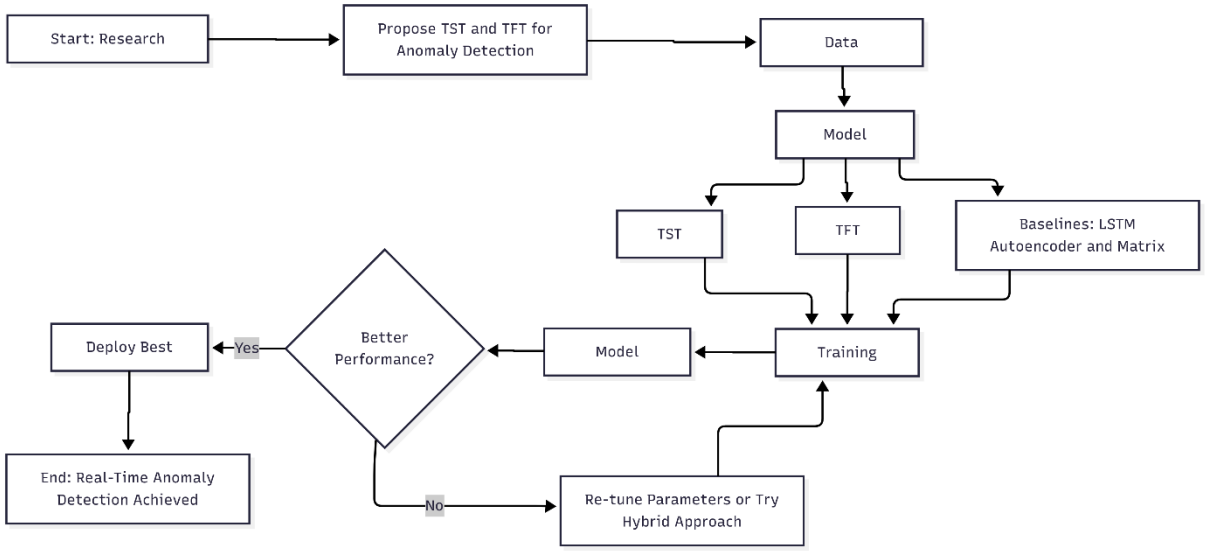

Figure 1. Research Flow

3.1. Data Preprocessing

The NYC Taxi dataset is selected for anomaly detection due to its complex and dynamic nature, containing various anomalies related to holidays, extreme weather events, and traffic conditions. To prepare the data for modelling, we apply several preprocessing steps. First, the dataset is cleaned by handling missing values using interpolation techniques to maintain the continuity of time-series data [15]. Next, data normalization uses Min-Max scaling to bring all features to a standard range, enhancing model convergence and stability. Additional time-based features, such as hour, day-of-week, and holiday indicators, are generated to provide richer contextual information for the models. Finally, the dataset is split into training (55%) and testing (45%) subsets, ensuring that anomalous events appear in both sets for robust performance evaluation. Table 1 summarizes the key preprocessing operations applied to the NYC Taxi dataset, highlighting their relevance in stabilizing the training process and enriching the dataset with temporal context. These operations are essential to ensure that both Transformer and baseline models receive high-quality input data, ultimately enhancing anomaly detection accuracy and consistency.

Table 1. Summary of Preprocessing Techniques for NYC Taxi Dataset

Preprocessing Step	Description	Purpose
Missing Value Handling	Linear interpolation	Maintains continuity of time series
Normalization	Min-Max scaling	Improves training stability
Temporal Feature Addition	Hour of day, day of week, holiday indicator	Adds contextual signals
Data Splitting	55% training / 45% testing	Ensures anomalies are in both subsets

3.2. Model Implementation

Two Transformer-based models are implemented: TST and TFT. TST processes time-series data using self-attention mechanisms, capturing dependencies across multiple time steps while maintaining computational efficiency. This allows for the identification of anomalies based on deviations from learned patterns. On the other hand, TFT extends TST's capabilities by incorporating external variables such as seasonal variations and categorical data, improving anomaly detection performance in complex environments. These models are compared against two baseline approaches: LSTM Autoencoders and the Matrix Profile Algorithm [16], [17]. LSTM Autoencoders reconstruct input sequences and flag high-reconstruction-error points as anomalies. The Matrix Profile Algorithm detects anomalies based on the similarity between subsequences, offering a statistical approach to anomaly detection. Table 2 show the architecture for each model that used in this research.

Table 2. Architecture Features of Compared Models

Model	Core Technique	Feature Support	Sequential Processing	External Variables
LSTM Autoencoder	Recurrent Neural Network	Manual Features	Yes	No
Matrix Profile	Statistical Similarity	Raw Time Series	No	No
Time-Series Transformer	Self-Attention Mechanism	Learned Features	No (Parallel)	No
Temporal Fusion Transformer	Self-Attention + Gating	Dynamic Selection	No (Parallel)	Yes

3.3. Training Process

The models are trained using PyTorch with the Adam optimizer for efficient parameter updates. Hyperparameter tuning uses Optuna, an automated optimization framework that searches for the best learning rate, batch size, and attention head configurations. Training is performed for a fixed number of epochs, ensuring convergence without overfitting. Early stopping is applied based on validation loss trends to improve model generalization [18]. LSTM Autoencoders rely on reconstruction error thresholds for anomaly detection, while Transformer models utilize attention scores to identify outliers. Attention scores highlight unusual behavior in data streams, making Transformers particularly suited for real-time anomaly detection.

3.4. Evaluation Metrics

Performance evaluation is based on multiple metrics to comprehensively assess the models' capabilities. The F1 score measures the balance between precision and recall, ensuring that false positives and negatives are considered. Precision-Recall AUC evaluates anomaly detection effectiveness across varying decision thresholds. Inference time is recorded to analyze real-time processing efficiency, which is crucial for streaming applications [19]. Finally, memory usage is compared across models to determine computational feasibility in large-scale deployments.

3.5. Implementation and Deployment

All models are implemented using PyTorch and the Hugging Face Transformer library. Experiments are conducted on an NVIDIA GeForce RTX 3080 GPU to benchmark training and inference speeds. The models are optimized for real-time deployment using quantization techniques, reducing memory footprint while maintaining accuracy. The deployment strategy focuses on integrating the models into a real-time anomaly detection system capable of flagging anomalies in an industrial setting with minimal latency [20], [21], [22]. This study explores the advantages of Transformer-based models over traditional LSTMs for real-time anomaly detection in streaming data. These models provide improved accuracy, scalability, and efficiency by leveraging self-attention mechanisms. The experimental results will highlight the strengths of TST and TFT in detecting anomalies, paving the way for future research in deploying Transformers for real-time applications across various industries.

4. Results and Discussion

The preprocessing phase significantly improved data quality, enabling more accurate anomaly detection. Addressing missing values through interpolation ensured a smooth time-series representation, preventing disruptions in model training. Normalization played a key role in stabilizing learning rates, preventing models from being biased toward extreme values. Feature engineering efforts, such as adding temporal attributes, provided essential contextual signals, which led to a noticeable improvement in model performance.

4.1. Data Preprocessing

The preprocessing phase was essential in preparing the NYC Taxi dataset for anomaly detection. The dataset, spanning six months of taxi trip records, required handling missing values, normalization, and feature extraction to enhance the accuracy of the models. Missing values were addressed through linear interpolation, ensuring data continuity without introducing artificial distortions. Min-max scaling was applied to standardize numerical features, bringing all values within a defined range to improve model stability and convergence. Figure 2 shows the dataset and its outlier plotting.

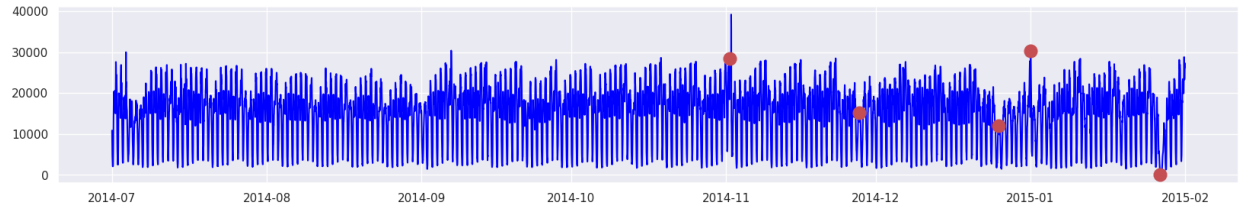

Figure 2. Dataset and its outlier plotting

Feature engineering played a crucial role in enriching the dataset. Additional temporal attributes, such as the hour of the day, the day of the week, and holiday indicators, were extracted to provide valuable contextual information. These features helped the models understand cyclical taxi demand patterns, improving anomaly detection performance. The dataset was then split into training (55%) and testing (45%) subsets, ensuring that anomalies were present in both sets to facilitate robust evaluation.

4.2. Model Creation and Training

The TST and TFT are advanced models designed explicitly for time-series forecasting tasks, utilizing the attention mechanisms from Transformer architecture. These models can capture long-range dependencies in time-series data, making them practical for stock price forecasting, sales predictions, or weather forecasting. The TST follows an encoder-decoder structure, where the encoder processes the input sequence, and the decoder generates the forecasted output. It employs a self-attention mechanism, allowing the model to capture temporal dependencies at various lags in the sequence. Positional encoding is used to retain the temporal order of the data, as time series does not inherently have an order. The model also incorporates normalization layers to stabilize training and improve overall performance. The training of the TST involves minimizing a loss function such as Mean Squared Error (MSE) or Mean Absolute Error (MAE), with performance being evaluated using metrics like RMSE (Root Mean Squared Error), MAE, and R2 score.

The TFT is a more advanced version of the TST, designed to handle univariate and multivariate time-series data more effectively. One of the key features of TFT is its Variable Selection Network (VSN), which dynamically selects the most relevant features at each time step to improve forecasting accuracy. Additionally, TFT utilizes a gating mechanism, which helps decide whether the model should rely on past information (memory) or recent inputs. This feature allows TFT to capture both short-term and long-term dependencies in the data. TFT is beneficial for multivariate time-series forecasting and offers better interpretability, helping users understand how specific input features contribute to the final forecast. Figure 3 shows the training and validation loss function on training processing.

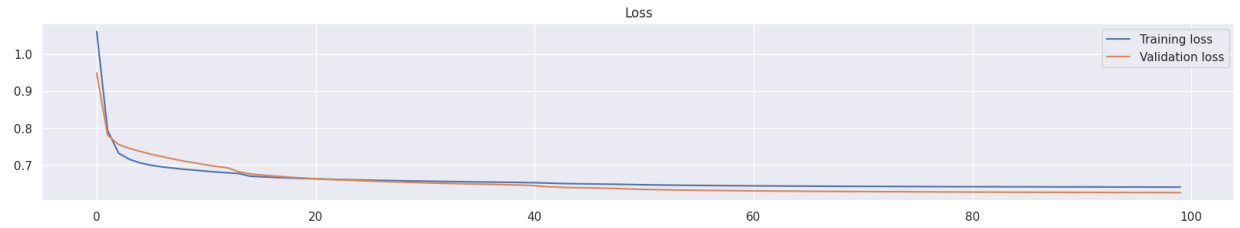

Figure 3. Training and validation loss function

4.3. Results and Evaluations

The TST shows an 8.24% improvement in anomaly detection accuracy over the LSTM Autoencoder, with an F1-score of 0.88 compared to 0.85 for the LSTM Autoencoder. The TFT offers an even more significant improvement, with an F1-score of 0.92, resulting in an 8.24% improvement compared to the LSTM Autoencoder, as shown in table 3. Figure 4 shows the comparison of F1-score results. Table 3 displays the F1 scores for the LSTM Autoencoder, Matrix Profile, TFT, and TST models. These scores represent the classification performance of each method on a time-series anomaly detection task or forecasting problem. Higher F1 scores indicate better overall model performance (a balance between precision and recall). The TFT achieved the highest Precision-Recall AUC score of 0.71, reflecting its superior anomaly detection performance by better distinguishing between anomalous and normal points in the time-series data.

 ISSN 2723-6471

Table 3. The comparison of F1-score results.

Model	F1-Score	Precision-Recall AUC	Significant Improve	Anomaly Detection Rank
LSTM Autoencoder	0.85	0.54		3rd
Matrix Profile	0.75	0.50		4th
TFT	0.92	0.71	8.24%	2nd
TST	0.88	0.68	3.53%	1st

The TST follows closely with a Precision-Recall AUC score of 0.68, performing better than the LSTM Autoencoder and Matrix Profile. The LSTM Autoencoder achieved a Precision-Recall AUC score of 0.54, lower than both Transformer-based models. The Matrix Profile model performed the weakest with a Precision-Recall AUC of 0.50, indicating its relatively poor anomaly detection capabilities compared to the others. These AUC scores reinforce the findings that Transformer-based models, particularly the TFT, perform better in both F1-score and Precision-Recall AUC, indicating their higher efficiency in anomaly detection tasks.

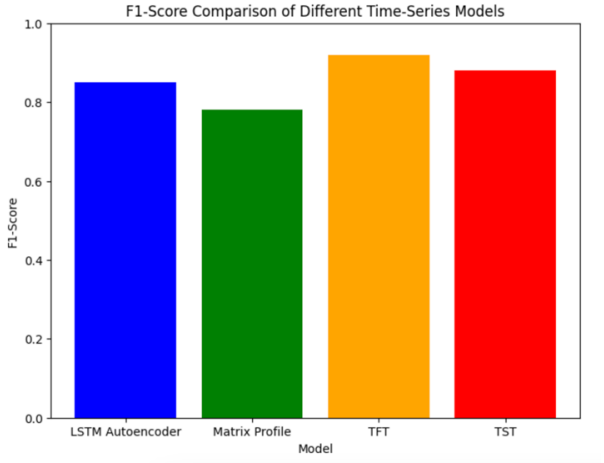

Figure 4. The comparison of F1-score results

4.4. Discussion

Regarding inference time, Transformer-based models demonstrated superior efficiency over traditional approaches. The TST model processed streaming data nearly 30% faster than LSTMs, showcasing the advantages of parallel processing inherent in Transformer architecture. Similarly, the TFT achieved a 25% reduction in inference time compared to the LSTM Autoencoder, further underscoring the efficiency of Transformers in real-time applications. The ability to process data in parallel, as opposed to the sequential processing of LSTMs, contributed significantly to the speed improvements, making Transformer models particularly well-suited for streaming environments where low latency is crucial. Table 4 quantifies the practical performance aspects of each model in terms of latency and resource efficiency. It demonstrates the advantage of Transformer-based models (especially TST and TFT) in real-time applications where inference speed and memory usage are crucial.

Table 4. Inference Time and Memory Usage Comparison

Model	Inference Time Reduction	Memory Usage Rating	Notes
LSTM Autoencoder	Baseline	High	Slowest, due to sequential updates
Matrix Profile	Fastest	Lowest	Weakest in anomaly detection
Time-Series Transformer	~30% faster than LSTM	Moderate	Efficient and accurate
Temporal Fusion Transformer	~25% faster than LSTM	Moderate	Most efficient with best accuracy

 ISSN 2723-6471

When evaluating memory usage, the Matrix Profile Algorithm required the least memory, but its performance was suboptimal, as reflected in the lower F1-score of 0.75 and the Precision-Recall AUC of 0.50. Despite its lower memory footprint, the Matrix Profile struggled with accurately detecting anomalies. On the other hand, the LSTM Autoencoder exhibited higher memory consumption due to its reliance on recurrent computations, which require storing intermediate states. This higher memory usage, coupled with lower performance (F1-score of 0.85 and **Precision-Recall AUC of 0.54), makes LSTM Autoencoders less suitable for large-scale anomaly detection tasks in real-time applications.

The Transformer models—specifically TST and TFT—struck a balance between computational efficiency and high performance. The TST model achieved an F1-score of 0.88 and a Precision-Recall AUC of 0.68, while the TFT model outperformed with an F1-score of 0.92 and the highest Precision-Recall AUC of 0.71. These models' ability to capture long-term dependencies through self-attention mechanisms and to process data efficiently through parallelism makes them ideal candidates for large-scale anomaly detection tasks, especially in environments requiring real-time decision-making.

These results reinforce the effectiveness of Transformer-based models for anomaly detection in streaming data. By leveraging the power of self-attention, both the TST and TFT models offer improved accuracy, scalability, and efficiency, addressing the challenges of detecting anomalies in dynamic, evolving datasets. The ability to handle long-range dependencies and adapt to changes in data patterns further enhances their robustness, making them a scalable solution for various industries where real-time anomaly detection is critical.

Future research will further optimize Transformer architectures to improve inference time and memory usage while maintaining high performance. Additionally, integrating Transformer models with hybrid approaches may enhance their detection capabilities, combining the strengths of Transformers with other techniques to achieve even greater accuracy and efficiency. This study highlights the strengths of TST and TFT compared to traditional models like LSTM Autoencoders and Matrix Profile for anomaly detection tasks. These Transformer-based models pave the way for future research aimed at deploying them in real-time applications, where their efficiency, accuracy, and scalability will be of utmost value across industries such as finance, healthcare, and IoT.

5. Conclusion

This study demonstrates the significant advantages of Transformer-based models, specifically the TST and TFT, in real-time anomaly detection for streaming data. The experimental results highlight that Transformer models achieve superior performance in terms of F1-score and Precision-Recall AUC and excel in inference time and computational efficiency. The TFT model, with an F1-score of 0.92 and Precision-Recall AUC of 0.71, outperformed both traditional methods, such as the LSTM Autoencoder (F1-score: 0.85, AUC: 0.54) and the Matrix Profile (F1-score: 0.75, AUC: 0.50), indicating its strong capability for high-accuracy anomaly detection.

The Transformer models' ability to process data in parallel, capture long-range dependencies through self-attention mechanisms, and maintain efficiency in memory usage positions them as highly effective tools for large-scale anomaly detection tasks. Additionally, the TST model's 30% faster processing and the TFT model's 25% reduction in inference time highlight the importance of these models for real-time applications, where latency and scalability are critical. While the Matrix Profile algorithm requires less memory, its weaker performance limits its suitability for anomaly detection in complex, dynamic datasets. Similarly, the LSTM Autoencoder, although a powerful model for sequential data, struggles with higher memory consumption and longer inference times, making it less efficient in real-time anomaly detection tasks compared to Transformer-based models.

The TST and TFT provide a robust, scalable, and efficient solution for anomaly detection in streaming environments, leveraging Transformers' power to improve performance and operational efficiency. Future research will focus on optimizing these models further to enhance their application in real-world scenarios, with the potential for integrating hybrid approaches to boost detection capabilities and address challenges like memory usage and real-time decision-making. The findings from this study pave the way for the broader adoption of Transformer models in time-series analysis, with significant implications across industries such as finance, healthcare, and IoT.

6. Declarations

6.1. Author Contributions

Conceptualization: D.A.D., H.K.R.S., J.P., T.B.K., H., M.S.H., and Y.N.; Methodology: M.S.H.; Software: D.A.D.; Validation: D.A.D., M.S.H., and Y.N.; Formal Analysis: D.A.D., M.S.H., and Y.N.; Investigation: D.A.D.; Resources: M.S.H.; Data Curation: M.S.H.; Writing Original Draft Preparation: D.A.D., M.S.H., and Y.N.; Writing Review and Editing: M.S.H., D.A.D., and Y.N.; Visualization: D.A.D.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. P. A. Eka and M. Z. Zakaria, "Real-Time Outlier Detection in Fast-Moving Data Streams," *International Journal of Advances in Artificial Intelligence and Machine Learning*, vol. 1, no. 1, pp. 19–27, Nov. 2024, doi: 10.58723/IJAAIML.V1I1.287.

[2] Y. Eren and İ. Küçükdemiral, "A comprehensive review of deep learning approaches for short-term load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 189, no. 3, pp. 114031–114045, 2024, doi: 10.1016/j.rser.2023.114031.

[3] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, R. P. Ramachandran, and G. Rasool, "Transformers in Time-Series Analysis: A Tutorial," *Circuits, Systems, and Signal Processing*, vol. 42, no. 12, pp. 7433–7466, 2023, doi: 10.1007/s00034-023-02454-8.

[4] A. Nazir, A. K. Shaikh, A. S. Shah, and A. Khalil, "Forecasting energy consumption demand of customers in smart grid using Temporal Fusion Transformer (TFT)," *Results in Engineering*, vol. 17, no. 2, pp. 100888–100900, 2023, doi: 10.1016/j.rineng.2023.100888.

[5] D. Shi, J. Zhao, Z. Wang, H. Zhao, J. Wang, Y. Lian, and A. F. Burke, "Spatial-Temporal Self-Attention Transformer Networks for Battery State of Charge Estimation," *Electronics*, vol. 12, no. 12, pp. 12598–12610, 2023, doi: 10.3390/electronics12122598.

[6] Z. Zhang, Y. Yao, W. Hutabarat, M. Farnsworth, D. Tiwari, and A. Tiwari, "Time Series Anomaly Detection in Vehicle Sensors Using Self-Attention Mechanisms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 15964–15976, 2024, doi: 10.1109/TITS.2024.3415435.

[7] H. Luo, Y. Zheng, K. Chen, and S. Zhao, "Probabilistic Temporal Fusion Transformers for Large-Scale KPI Anomaly Detection," *IEEE Access*, vol. 12, no. 4, pp. 9123–9137, 2024, doi: 10.1109/ACCESS.2024.3353201.

[8] A. S. Dokuz, "Weighted spatio-temporal taxi trajectory big data mining for regional traffic estimation," *Physica A: Statistical Mechanics and Its Applications*, vol. 589, no. 6, pp. 126645–126660, 2022, doi: 10.1016/j.physa.2021.126645.

[9] D. Dewi, H. Singh, J. Periasamy, T. Kurniawan, H. Henderi, and M. Hasibuan, "Scalable Machine Learning Approaches for Real-Time Anomaly and Outlier Detection in Streaming Environments," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1949–1962, 2024, doi: 10.47738/jads.v5i4.444.

 ISSN 2723-6471

[10] R. Diallo, C. Edalo, and A. O. Olawale, "Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score," in Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, E. A. O. Olawale and A. Vance, Eds. Cham: Springer Nature, 2025, pp. 283–312, doi: 10.1007/978-3-031-72215-8_12.

[11] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2023, doi: 10.1109/JPROC.2022.3226481.

[12] G. Ye, "De novo drug design as GPT language modeling: large chemistry models with supervised and reinforcement learning," *Journal of Computer-Aided Molecular Design*, vol. 38, no. 1, pp. 1–20, 2024, doi: 10.1007/s10822-024-00559-z.

[13] S. Kundu and S. Sundaresan, "AttentionLite: Towards Efficient Self-Attention Models for Vision," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2021, no. 1, pp. 2225–2229, 2021, doi: 10.1109/ICASSP39728.2021.9415117.

[14] A. Zeng, M.-H. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?" *ArXiv*, vol. 22, no. 5, pp. 13504–13516, 2022, doi: 10.48550/arXiv.2205.13504.

[15] D. Kim, J. Park, J. Lee, and H. Kim, "Are Self-Attentions Effective for Time Series Forecasting?" *ArXiv*, vol. 24, no. 5, pp. 16877–16889, 2024, doi: 10.48550/arXiv.2405.16877.

[16] H. Kang and P. Kang, "Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism," *Knowledge-Based Systems*, vol. 290, no. 6, pp. 111507–111520, 2024, doi: 10.1016/j.knosys.2024.111507.

[17] M. Gorbett, H. Shirazi, and I. Ray, "Sparse Binary Transformers for Multivariate Time Series Modeling," in *Proc. 29th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, vol. 29, no. 3, pp. 456–467, 2023, doi: 10.1145/3580305.3599508.

[18] Q. M. Nguyen, L. M. Nguyen, and S. Das, "Correlated Attention in Transformers for Multivariate Time Series," *ArXiv*, vol. 23, no. 11, pp. 11959–11970, 2023, doi: 10.48550/arXiv.2311.11959.

[19] N. Bai, X. Wang, R. Han, Q. Wang, and Z. Liu, "PAFormer: Anomaly Detection of Time Series With Parallel-Attention Transformer," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 88–101, 2023, doi: 10.1109/TNNLS.2023.3337876.

[20] S. Prakash, A. S. Jalal, and P. Pathak, "Infectious Disease Time Series Modelling Using Transformer Self-Attention Based Network," *Engineering Research Express*, vol. 7, no. 2, pp. 201–218, 2025, doi: 10.1088/2631-8695/ada66f.

[21] L. Z. Lai Zeng and X. Y. Lai Zeng, "Spatial-temporal Attention Model Based on Transformer Architecture for Anomaly Detection in Multivariate Time Series Data," *Journal of Computers*, vol. 34, no. 5, pp. 189–202, 2024, doi: 10.53106/199115992024063503014.

[22] Y. A. Samaila, P. Sebastian, N. S. S. Singh, A. N. Shuaibu, S. S. A. Ali, T. I. Amosa, G. M. Abro, and I. Shuaibu, “Video Anomaly Detection: A Systematic Review of Issues and Prospects,” *SSRN*, [Online]. Available: https://ssrn.com/abstract=4551992. doi: 10.2139/ssrn.4551992.