Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

An adaptive transformer model for anomaly detection in wireless sensor networks in real-time

A. Siva Kumar [a,*], S. Raja [b], N. Pritha [c], Havaldar Raviraj [d], R. Babitha Lincy [e], J. Jency Rubia [f]

[a] Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India
[b] School of Electrical & Communication, Department of Electronics & Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Tamilnadu, India
[c] Department of ECE, Panimalar Engineering College, Tamilnadu, India
[d] Department of BME, KLE Dr. M S Sheshgiri College of Engineering and Technology, Belgaum, India
[e] Computer and Communication Engineering Department, Sri Eshwar College of Engineering, Tamilnadu, India
[f] Department of ECE, K. Ramakrishnan College of Engineering, Tamilnadu, India

ARTICLE INFO

Keywords:
Anomaly detection
WSN
Sensor
Transformer
Spatio-temporal attention
Adaptive

ABSTRACT

Defense, environmental monitoring, healthcare, home automation, and other fields are just a few of the many that make use of wireless sensor networks. There are sensor nodes in these networks that are sent out into the field to collect data and relay it back to the network's hub. The data streams may be infused with anomalous data due to interferences, malfunctioning nodes, etc. The security of the network depends on the prompt identification of any irregularities. Due to noise, missing data, and the dynamic nature of the network, identifying and detecting abnormal data is difficult.This study introduces a dynamic context-capturing deep learning model for sensor data anomaly detection, constructed as a Transformer with a spatio-temporal attention mechanism. The proposed model is trained and tested with a public dataset built from real-time sensor data captured in a water treatment plant. With an accuracy of 0.09142, F1 of 0.9109, and precision of 0.9191, the STA-Tran model outperforms the best existing methods.

1. Introduction

Time-series data captured by sensors in Wireless Sensor Networks (WSNs) are prone to a number of different sources of noise and anomalies. The data are skewed, as they are generated by sensor nodes with different energy levels. Detection of anomalous data is important in commercial, industrial, healthcare and military applications as they may interfere with the workflow of the critical processes or decisions. Automated anomaly detection models are useful for extracting information from time-series data captured by sensors in WSNs [1]. Designing these models is challenging due to the diversities in the number of devices, data types, data lengths, sensor placement, and power supplies.

For anomaly detection, probabilistic models like Hidden Markov Models (HMMs) [2] and Bayesian Networks (BNs) [3] have seen extensive application. HMMs are based on Markov Chain, which are not suitable for continuous-time data, and BNs are commonly used to model non-linear time-series data. However, the learning of BNs is not straightforward, and theyalsosuffer from the need of prior knowledge of the data distribution.

Anomaly detection often makes use of machine learning models like Support Vector Machines (SVMs) [4]. However, SVMs require feature vectors for the training data. Thus, feature extraction is required for the data prior to the training process. The feature vectors are often obtained from the data pre-processing and feature selection approaches. For the same data, different feature vectors may be obtained, leading to different classification results.

Anomaly detection in time series data has become more popular as a result of the development of deep learning, with models based on RNNs [5,6] and LSTMs [5,6] finding widespread use. These models can both acquire knowledge from historical information and apply that knowledge to brand new situations. It is unclear if the data should be pre-processed for the anomaly detection, which drives up the cost of using such models. However, training such models requires a huge

[*] Corresponding author.
E-mail addresses: siva.aaru53@gmail.com (A. Siva Kumar), srajaphd2011@gmail.com (S. Raja), prithabe28@gmail.com (N. Pritha), dr.rhhavaldar@klescet.ac.in (H. Raviraj), rblincy@gmail.com (R. Babitha Lincy), jencyrubia@gmail.com (J. Jency Rubia).

https://doi.org/10.1016/j.measen.2022.100625
Received 2 October 2022; Received in revised form 17 November 2022; Accepted 2 December 2022
Available online 10 December 2022

quantity of data, which is often acquired by pre-processing. Further, these models are prone to vanishing and exploding gradient problems.

Transformer models have recently found utility in Natural Language Processing (NLP), sequence modelling, computer vision, and voice recognition applications [7,8] because of their ability to capture long-term relationships in the data and to manage extended sequences without pre-processing.

This research uses the potential of Transformer networks and a spatio-temporal attention mechanism to identify outliers in time-series sensor data, satisfying the need for a robust anomaly detection model for sensor data. The adaptive transformer model requires little input data during training and automatically adjusts to fresh input without any human intervention. The model is resistant to the amount of the training data and can parse large sequences directly.

The contributions of this research are as below.

1 For the first time, an adaptive Spatio-Temporal Attention Transformer (STA-Tran) model is suggested for detecting anomalies in time-series sensor data for WSNs.
2 After training and testing on a publicly available dataset, the model outperforms its representative baselines.
3 STA-Tran can be extended to time-series classification, regression, and clustering problems in diverse applications such as financial modelling, health monitoring and smart homes.

This paper's structure continues below. The second part of this paper provides an overview of the deep learning models used for anomaly identification in time-series data in recent years. In Section 3, we will go through the data used and the methodology behind this study. Section 4 describes the suggested model, while Section 5 provides empirical data and comparative analyses. Portion 6 is the last section of the paper.

2. Realted work

In this part, we will discuss some of the key works that have informed our study. In their exhaustive study of deep learning models for anomaly identification, Choi et al. [9] describe the many sorts of anomalies that may be found in time-series data, such as point anomaly, context anomaly, and collective anomaly.Identifying the limitations of conventional models such as lack of failure modes and complexity of sensor data, the authors advocate the usage of deep learning models owing to their ability to model and detect complex nonlinear relationships among the data.

In time-series analysis, temporal correlation of the data is a very important aspect, which is often not considered in the training of conventional anomaly detection models. RNNs are most popular deep learning models for time-series data. They are feedforward neural networks, in which the input data are processed one time step at a time. The network produces results based on the input it received at the prior time step. RNNs are well-suited to modelling time-series data and other types of sequential information [10,12].

Long short-term memory (LSTM) RNNs model and learn temporal relationships in data sets over extended periods of time. They have been employed for anomaly identification in time-series data and can capture long-term temporal relationships. Anomaly detection in spacecraft telemetry is performed using a model called LSTM-NPT [11,13] that is based on LSTM and Nonparametric Dynamic Thresholding (NPT). In this model, LSTM is used in predicting the data and NPT is applied on this data for unsupervised dynamic anomaly detection. If you use data from the Mars Science Laboratory (MSL), this model has the best accuracy at 92.6%. For anomaly identification in multi-variate time-series data, another LSTM-based model, the Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) [14], is presented.

To identify outliers in high-dimensional data, the Deep Autoencoding Gaussian Mixture Model (DAGMM) [15] uses an unsupervised learning approach. The input is encoded with minimal dimensionality using an Auto Encoder (AE), and then reconstructed. A Gaussian Mixture Model (GMM) is used to estimate density and spot outliers based on the reconstruction error. The model is jointly optimized for end-end training. This model achieves >14% improvement in F1 values compared to the standard value. However, DAGMM does not explicitly model the data distribution and hence, cannot be applied to cases where the data distribution is not known.

To model time-series data with context, the Multivariate Anomaly Detection with GAN (MAD-GAN) [16] framework uses Generative Adversarial Networks (GANs) with Long Short-Term Memory (LSTM) backbones. This model does not consider the data stream for processing, and it operates on a set of variables to model the interactions between them. This model performs discrimination based anomaly detection by training the discriminator to detect anomalous data. It also performs anomaly detection by training the generator to construct realistic samples and measuring the reconstruction error. This model is tested on the Secure Water Treatment (SWaT) [17] dataset, achieving a precision of 99.99% and Recall value of 54.80%. However, the model may become instable due to complex interactions between variables.

Merlin [18] is an algorithm for detecting abnormalities of arbitrary sizes called 'discords' from real-time data. This algorithm is based on the prior definition of a threshold value to identify discords from the subsequences of multi-variate time-series. This cutoff is less than the discord distance, which is the space between a throwaway and its nearest neighbour.This algorithm is capable of identifying very subtle anomalies from standard datasets. However, it fails with inappropriate threshold definition.

The UnSupervised Anomaly Detection (USAD) [19] is an adversely trained AE model for anomaly detection from multi-variate time-series. This framework first divides the data into intervals, then it trains a model to identify typical data and last it awards an anomaly score to each projected data point. The model learns to detect anomalies by minimizing this score. This model achieve a highest F1 score of 0.7875 on theSWaTdataset. However, it does not consider the correlation between the time frames.

Multivariate time-series anomaly detection using a graph attention network was suggested by Zhao et al. [20], who recognized the need of recording the interrelationship of the different time periods (MTAD-GAT).This model treats each univariate time-series component as a feature and captures the semantic and temporal contexts between them, using two graph attention layers. These layers model the relationship between the nodes in a graph, where each node is a feature or time stamp. Based on the MSL data, this model gets an F1 score of 0.9084. However, it's possible that the model's efficacy shifts depending on the duration of the time periods.

A similar work in this context proposed by Deng and Hooi [21] proposes Graph Deviation Network (GDN) to learn the relationships between sensors to detect anomalies. It models a group of sensors as a graph with edges representing their relationships. This model forecasts the behaviour of a sensor using an attention mechanism which learns from the current scenario depicted by the graph. This model detects anomalies by computing the anomaly score for each node and identifying the deviating one, and achieves an F1 score of 0.81 on the SWaT dataset.

The Convolutional Autoencoding Memory network (CAE-M) [22] employs two subnetworks, one each to define the special dependencies and the other to define temporal dependencies between the data. These modules are constructed using a deep convolutional AE and a bidirectional LSTM equipped with an attention mechanism. As a result, the CAE-M optimizes the subnetworks jointly for anomaly detection on multi-sensor data using a memory network that captures these dependencies concurrently. This model achieves a mean F1 of 0.9961 on the Cyclic Alternating Pattern (CAP) [23] dataset of physiological signals.

The pre-processing, feature image creation, and anomaly detection stages make up the Attention-Based Convolutional Long Short-Term

Table 1
Training and testing dataset description.

No. of Training Data	Number of Testing Data	No. of Sensors	Anomaly %
496800	449919	51	11.98

Memory (ConvLSTM) Autoencoder with Dynamic Thresholding (ACLAE-DT) [24] framework. Initially, the raw time-series is normalized and transformed into multi-variate time-series by sliding a window over the normalized data. In order to rebuild a time series, an AE based on convolutional LSTM is fed feature pictures that depict the dependencies between various pairs of time series. As part of its training, the AE is asked to recreate the feature photos with as little mistake as possible. Using a dynamic thresholding technique, outliers, causes of anomalies, and origins of anomalies are isolated from rebuilt feature pictures to discover abnormalities. This model is evaluated with the CNC Mill Tool wear data [25] achieving precision, recall and F1 values of 0.95, 0.88 and 0.92 respectively. However, this model may result in sub-optimal performances when the AE does not reconstruct the feature images perfectly.

Understanding the significance of the attention based models in several problems, recently researchers have resorted to vanilla Transformer models due to their inbuilt attention mechanisms capable of capturing long-term dependencies between data. This model is widely employed in NLP, Machine Translation (MT), image classification and segmentation tasks. This model has not so far been used in the computer and machine vision problems involving time-series data.

First employed in anomaly detection of multi-variate time-series data, TranAD [26] is a Transformer model. This model follows a two phase adversarial training approach so that it generalizes well, for anomaly detection on arbitrary data sequences. In the first phase, the model reconstructs the input sequences and the reconstruction error called the focus score is used in the second phase to extract short temporal trends called self conditioned outputs from the regions of high deviations. This model is tested on several publicly available datasets and outperforms prior works. For the SWaT dataset, this model achieves an F1 score of 0.8151. However, this model achieves F1 value of 0.9780 on the MIT-BIH Supraventricular Arrhythmia Database (MBA) [27] dataset.

This review shows that Transformer models are better than the conventional models which need additional attention mechanisms. This model has not yet been widely used in anomaly detection, and investigation in this context can lead to prospective tools for detection, localization and classification of anomalies in several domains employing WSNs for data collection.

3. Materials and methods

In this part, we will discuss the methodologies that underpinned this study and the dataset that was utilized to train and evaluate STA-Tran.

3.1. Dataset and details of implementation

The STA-Tran is trained and tested with the SWaT dataset which has 6 stages. Staring from the first stage which regulates the inflow of water to the last stage in which the membranes of the ultrafiltration units are cleaned, this infrastructure comprises several sensors, actuators, piping, purification mechanisms. The sensor data such as water level, flow rate, pressure etc. are captured for analyses. This dataset comprises sensor data collected by normal operation of the plant for 7 days and abnormal operation for 4 days. The data for abnormal operation are collected by launching a set of attacks. The dataset is annotated by experts according to the system behaviours and it is a benchmark for several anomaly detection systems.The data set's characteristics are shown in Table 1.

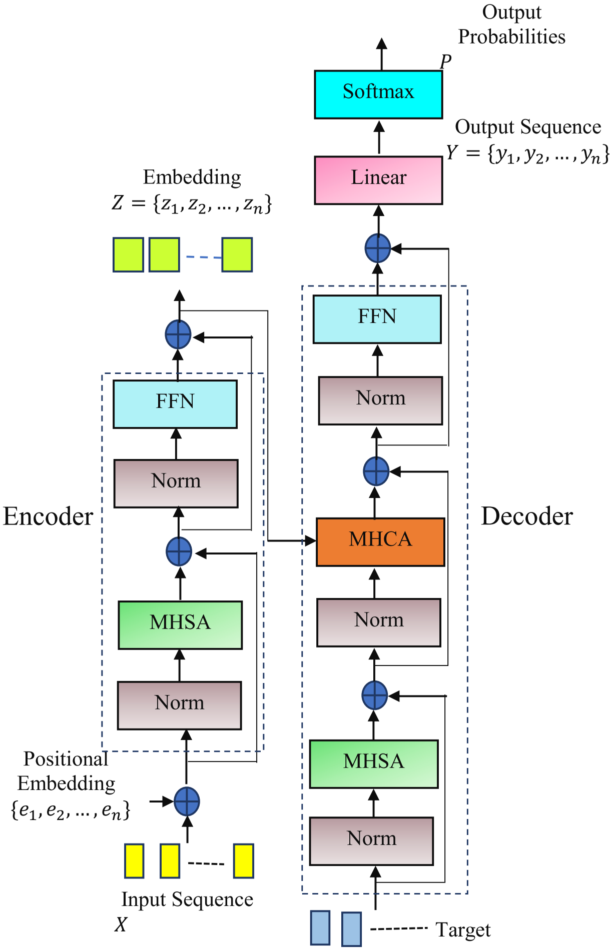

Fig. 1. Transformer architecture.

3.2. Transformer Architecture

As can be seen in Fig. 1, the Transformer employs an encoder-decoder architecture with three types of attention: the self-attention of the input-sequence, the attention of the encoder, and the attention of the decoder.

The encoder consists of a multi-head self-attention mechanism (MHSA) on top of a position-wise completely linked feed-forward network (FFN).Given an input sequence $X = \{x_1, x_2, \dots, x_n\}$, the encoder generates a latent representation $Z = \{z_1, z_2, \dots, z_n\}$ by transforming X with the MHSA and FFN. The sequence Z is transformed by the decoder to generate the output sequence $Y = \{y_1, y_2, \dots, y_n\}$ which carries the contextual information of X.

The MHSA relates the positions of the elements x_i to compute the attention. Along with each element x_i, its position e_i is concatenated and given as input to the MHSA. Layer Normalization (LN) is applied on each $\{x_i, e_i\}$ to enhance the model's capability of learning effectively from unstructured inputs. In addition, skip connections are used between the components of the encoder and decoder for residual learning. The normalized form of each $\{x_i, e_i\}$ is transformed into three vectors namely the Query Q, Key K and Value V employing three weight matrices W^Q, W^K and W^Q as in equations (1)–(3).

$$Q = \{x_i, e_i\}W^Q \quad (1)$$

$$K = \{x_i, e_i\}W^k \quad (2)$$

$$V = \{x_i, e_i\}W^V \quad (3)$$

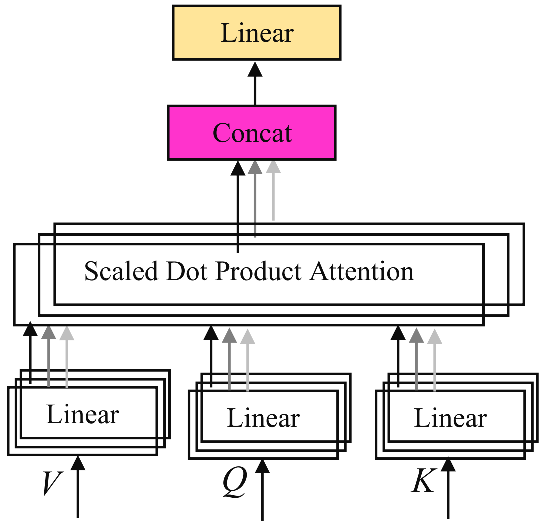

Fig. 2. MHSA mechanism.

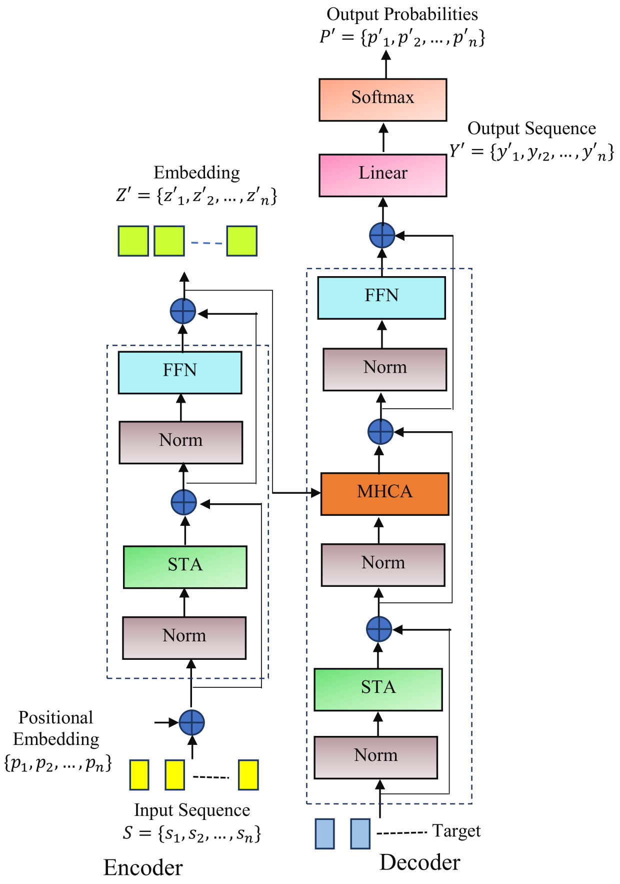
Output Probabilities
$P' = \{p'_1, p'_2, \dots, p'_n\}$
Softmax
Output Sequence
$Y' = \{y'_1, y_{\prime 2}, \dots, y'_n\}$
Linear
Embedding
$Z' = \{z'_1, z'_2, \dots, z'_n\}$
FFN
Norm
MHCA
Norm
STA
Norm
FFN
Norm
STA
Norm
Positional Embedding
$\{p_1, p_2, \dots, p_n\}$
Input Sequence
$S = \{s_1, s_2, \dots, s_n\}$
Target
Encoder
Decoder

Fig. 3. STA-Tran architecture.

For each element, the self attention score is computed from these vectors as in equation (4) where d_k is the dimension of K_i.

$$AS(Q_i, V_i, K_i) = softmax\frac{(Q_iK_i^T)}{\sqrt{d_k}}V_i \tag{4}$$

The schematic of the MHSA mechanism is shown in Fig. 2 which performs attention computation in parallel on multiple elements.

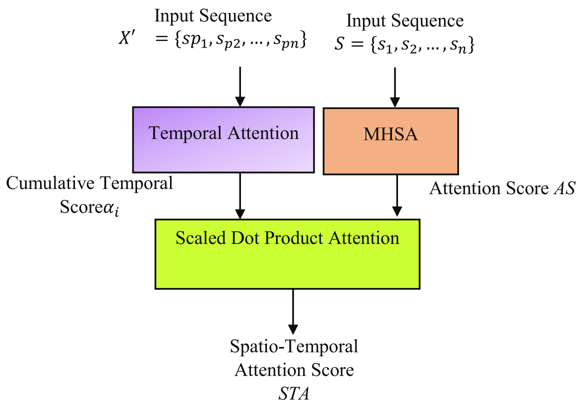
Input Sequence
$X' = \{sp_1, s_{p2}, \dots, s_{pn}\}$
Input Sequence
$S = \{s_1, s_2, \dots, s_n\}$
Temporal Attention
MHSA
Cumulative Temporal Scoreα_i
Attention Score AS
Scaled Dot Product Attention
Spatio-Temporal Attention Score
STA

Fig. 4. Spatio-temporal attention.

4. Proposed STA-tran network

As can be seen in Fig. 3, the STA-Tran network for anomaly detection is realized as a Transformer network with Spatio-Temporal Attention mechanisms (STA) in both the encoder and the decoder. At first, the time series is transformed into n frames of fixed length, where s is the number of samples in the sequence of sensor data F = s-1, s-2, s-N. Each data frame is normalized and processed in the same way sentences are processed by the Transformers. For each frame F_i, N positional embeddings are constructed corresponding to each sensor data as $P_i = \{p_1, p_2, \cdots, p_N\}$. Each data is combined with the respective embedding to form a pair $\{s_i, p_i\}$.

A sequence X' of these pairs is given as input to the STAmechanism shown in Fig. 4 which computes the attention in the spatial and temporal dimensions and combines them. The spatial attention is computed applying (4) as in equation (5) where the triple (Q_i', V_i', K_i') is the Query, Value and Key values constructed from each $\{s_i, p_i\}$.

$$AS(Q_i', V_i', K_i') = softmax\frac{(Q_i'K_i'^T)}{\sqrt{d_k}}V_i' \tag{5}$$

The temporal attention is computed as in (6) where w_i is the learned weight vector and t_j is the j^{th} time-step of the ith data frame.

$$\alpha_{ij} = \frac{exp(w_i \cdot t_j)}{\sum_{j=1}^{N} exp(w_i \cdot t_j)} \tag{6}$$

The spatio-temporal attention for each element s_i is computed by combining (5) and (6) in (7) where α_i is the cumulative temporal attention for s_i.

$$Attention(s_i) = softmax\frac{(AS(Q_i', V_i', K_i'))}{\sqrt{d_k}}\alpha_i \tag{7}$$

The encoder produces the latent representation of X' from the attention scores as $Z' = \{z_1', z_2', \dots, z_N'\}$.

Similarly, the decoder learns attention from the encoder input and the query embeddings to generate the output sequence $Y' = \{y_1', y_2', \dots, y_N'\}$ with the probabilistic scores.

The STA-Tran is trained with both normal and anomalous samples to minimize the loss as in (8), where y_i' is the predicted output for x_i', and $(P(y_i'|X'))$ is the predicted probability for each element y_i' in X' as in (9).

$$L(X', Y') = \sum_{i=1}^{N} logP(y_i'|X') \tag{8}$$

$$\left(P\left(y_i'\middle|X'\right)\right) = softmax\frac{exp(y_i'\alpha_i)}{\sum_{i=1}^{N} exp(y_i'\alpha_i)} \tag{9}$$

5. Experimental results and discussions

In this part, we provide the testing dataset evaluation findings and state-of-the-art model comparisons from our STA-Tran evaluation.

5.1. Experimental setup

The training dataset is split in the ratio 80:20 into training and validation subsets to evaluate STA-Tran. MATLAB running on Windows 10 64-bit with an Intel(R) Core(TM) i5-2400 CPU at 3.10 GHz is used to construct and test this model. Model training is terminated when the validation loss reaches 1E-3, which typically occurs after 500 epochs. The Adam optimizer is used to fine-tune the model, with an initial learning rate of 0.001 and subsequent reductions of a factor of 10 after every 50 epochs.

5.2. Evaluation metrics

Generally, the anomaly detection models are evaluated with the accuracy, specificity, sensitivity (recall), precision and F1 measures defined in equations 10–14. From the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values obtained while analysing STA-Tran with the testing dataset, we calculate these measures.

Accuracy measures how well a model can separate out typical data points from outliers in a test dataset, as in (10).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{10}$$

The sensitivity of a test is defined as the fraction of normal instances that were accurately recognized. As in (11).

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{11}$$

The model's specificity may be defined as its propensity to accurately identify the test dataset's anomalous instances out of the total number of test dataset anomalous cases, as in (12).

$$Specificity = \frac{TN}{(TN + FP)} \tag{12}$$

Precision is the number of normal instances identified correctly out of the total number of normal cases predicted by the model as in equation (13).

$$Precision = \frac{TP}{(TP + FP)} \tag{13}$$

The model's robustness to unbalanced classes in the dataset may be quantified by the F1 score. The evaluation is carried out using the recall and precision indicators, as in (14) above.

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{14}$$

The above metrics range from 0 to 1, signifying perfect detection of the normal and abnormal cases when the values are closer to 1.

The Wilcoxon pairwise signed rank test is used to assess a classifier model's Area Under Curve (AUC). The discriminatory power of the model between the two classes is quantified by the area under the curve (AUC). The poorest classifier would have an AUC of 0, a random classifier would have an AUC of 0.5, and a perfect classifier would have an AUC of 1.0.

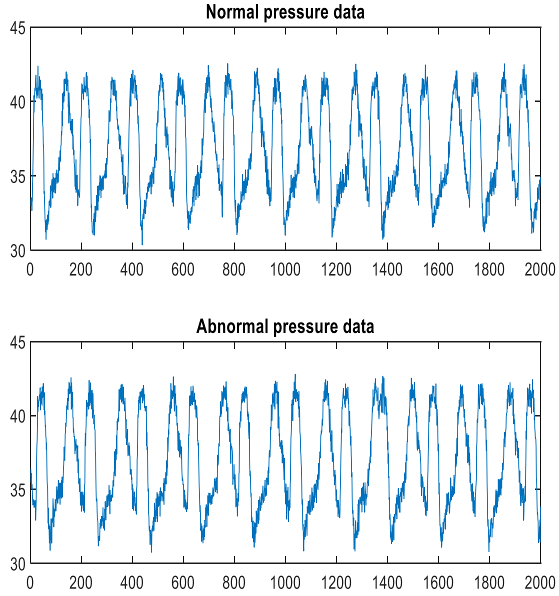

Fig. 5. Normal and anomalous sensor data.

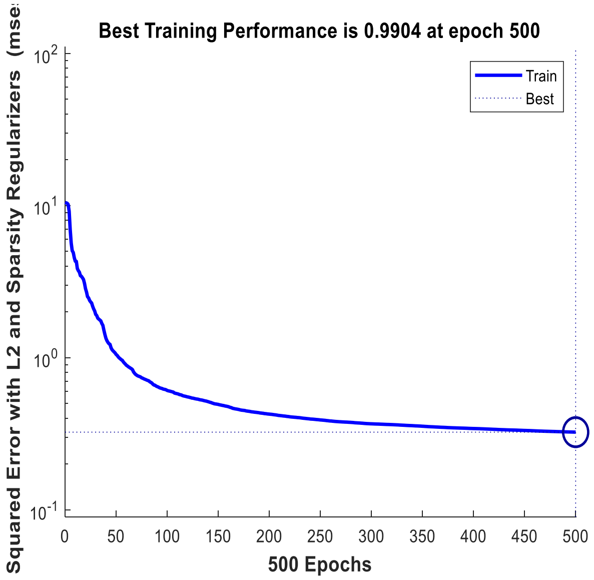

Fig. 6. Training performance of STA-Tran.

5.3. Anomaly detection results

The STA-Tran proposed model is evaluated with the testing data which is the pressure data captured by the sensors. The normal and abnormal data captured from the sensors is shown in Fig. 5.

Fig. 6 displays the results of training the STA-Tran on the training dataset. Multiple frames of the same size are formed using the provided time-series data, and then positional embeddings are generated for each frame. The sequence containing the data concatenated with the embeddings is given as input to the encoder and the output sequence is generated at the decoder. The output consists of the output sequence with the corresponding probabilities of the two target class, "normal" and "anomalous". The loss between the true and predicted classes is used to measure the model's efficacy. At the 500th epoch, the best training accuracy of 0.9904 is observed for the model. At about 400 epochs, the

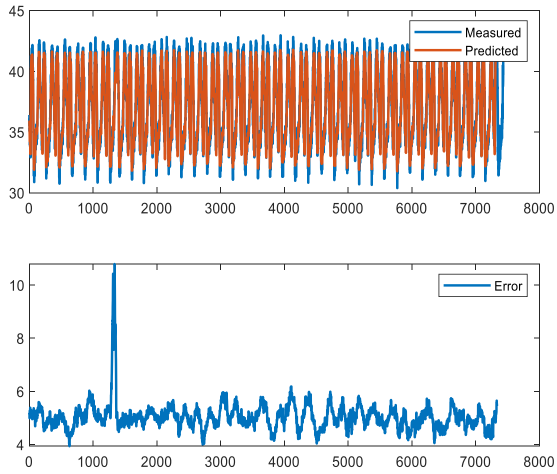

Fig. 7. Actual data and STA-Tran predicted data.

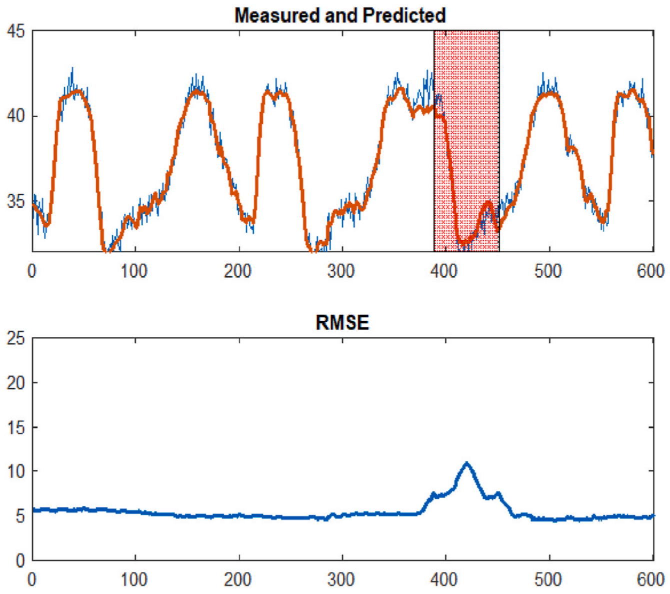

Fig. 8. Anomaly detection by STA

model seems to settle as the training loss decreases seemingly linearly with time.

After training, the model is tested with the entire test dataset to predict the anomalous samples. Similar to training data, the test dataset is split into frames and positional embeddings are created. This sequence called the target sequence is given as input to the first layer of the decoder. The decoder combines the output of the STA at this layer with that of the latent representation constructed by the encoder at the MHCA. The decoder generates the target sequence with the probabilities of target classes for each data in the frame. Fig. 7 shows the actual data in a frame used to train STA-Tran and the target sequence constructed by the decoder. Differences between the actual and expected data in each frame are represented by the Root Mean Squared Error (RMSE) of the prediction.

In addition to the above, the anomalous data is detected by the model by comparing the differences between the actual and predicted data model over small windows as shown in Fig. 8. It is seen that STA-Tran is capable of predicting the anomalies over small windows without ambiguities. This figure visualizes the error over the window presenting a visualizing the behaviour of the model. By setting the error thresholds, the model can be trained to capture subtle differences between the actual and predicted data for mission-critical applications.

Table 2 compares the proposed model to state-of-the-art approaches in terms of performance metrics for anomaly detection. Best values are shown in red letters, with the next best values shown in blue fonts. The best performance is attained by the planned STA-Tran, followed by TranAD.

An in-depth analysis of the results shows that attention based models outperform other deep learning based models. It is obvious that attention based models are designed to capture the most discriminative data and thus they achieve better results than the deep learning based models.However, Transformer models with built-in attention mechanisms are very efficient in capturing long-range dependencies while processing large-scale data. The transformer model is best suited for anomaly detection in sensor data as the sensors measure data at different time intervals and at the same time data is collected from multiple locations. The efficacy of the transformers in learning the target classes from multi-variate time-series data is testified by the anomaly detection results achieved by TranAD and the proposed STA-Tran. From Table 2, it is seen that STA-Tran shows considerable performance gain compared to TranAD which is attributed to the STA mechanism which captures the context at both the spatial and temporal channels.

6. Conclusion

This work introduces STssA-Tran, a revolutionary deep learning Transformer model for automating anomaly detection in WSNs' multivariate sensor data. A standard Transformer serves as the basis for this model, which has been improved with a STA mechanism that can adjust to changing conditions. This model captures the context of the input sequences in the spatial and temporal dimensions for discerning the normal and anomalous data. This model is tested on the SWaT dataset, and it outperforms other deep learning models achieving the highest classification accuracy of 0.9142. This model is a baseline for future models for intrusion detection, anomaly localization, detection and

Table 2
Comparison of anomaly detection performance with state-of-the-art models.

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
STA-Tran	**0.9142**	**0.9172**	**0.9158**	**0.9191**	**0.9109**	**0.9115**
TranAD Tuli et al. [26] (2022)	**0.8453**	**0.8341**	**0.8632**	**0.8491**	**0.8455**	**0.8788**
ACLAE-DT [24] Tayeh et al. (2022)	0.8291	0.8287	0.8258	0.8177	0.8193	0.8203
CAE-M Zhang et al. [22] (2021)	0.8164	0.8181	0.8054	0.8022	0.8188	0.8201
GDN Deng and Hooi [21] (2021)	0.8084	0.8077	0.7856	0.8121	0.8086	0.8406
USAD Audibert et al. [19] (2020)	0.7169	0.7061	0.7343	0.7206	0.7171	0.7494
MTAD-GAT Zhao et al. [20] (2020)	0.8001	0.8096	0.8067	0.8136	0.8123	0.8113
Merlin Nakamura et al. [18] (2020)	0.8053	0.8154	0.8111	0.8086	0.8154	0.8149
MSCRED Zhang et al. [14] (2019)	0.8113	0.8005	0.8150	0.8149	0.8115	0.8036
MAD-GAN Li et al. [16] (2019)	0.8267	0.8169	0.8423	0.8300	0.8268	0.8560
LSTM-NDT Hundman et al. [13] (2018)	0.8323	0.8323	0.8182	0.8356	0.8224	0.8221
DAGMM Zong et al. [15] (2018)	0.8086	0.7990	0.8239	0.8119	0.8088	0.8373

classification at different levels of granularities.STA-Tran can be extended to clinical and industrial settings which capture multi-sensor data, in which identification of anomalous data is a non-trivial problem.

CRediT authorship contribution statement

A. Siva Kumar: Conceptualization, Data curation, Investigation, Visualization, Writing – review & editing, Writing – original draft. **S. Raja:** Conceptualization, Visualization, Writing – review & editing, Writing – original draft. **N. Pritha:** Formal analysis, Writing. **Havaldar Raviraj:** Formal analysis, Writing – review & editing. **R. Babitha Lincy:** Formal analysis, Writing – review & editing. **J. Jency Rubia:** Conceptualization, Formal analysis, Supervision.

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.

Data availability

Data will be made available on request.

References

[1] A. Blázquez-García, A. Conde, U. Mori, J.A. Lozano, A review on outlier/anomaly detection in time-series data, ACM Comput. Surv. 54 (3) (2021) 1–33.

[2] W. Wang, Q. Chen, X. He, L. Tang, Cooperative anomaly detection with transfer learning-based hidden Markov model in virtualized network slicing, IEEE Commun. Lett. 23 (9) (2019) 1534–1537.

[3] M. Injadat, F. Salo, A.B. Nassif, A. Essex, A. Shami, Bayesian optimization with machine learning algorithms towards anomaly detection, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, December, pp. 1–6.

[4] Q. Ma, C. Sun, B. Cui, X. Jin, A novel model for anomaly detection in network traffic based on kernel support vector machine, Comput. Secur. 104 (2021), 102215.

[5] M. Canizo, I. Triguero, A. Conde, E. Onieva, Multi-head CNN–RNN for multi-time-series anomaly detection: an industrial case study, Neurocomputing 363 (2019) 246–260.

[6] D. Wu, Z. Jiang, X. Xie, X. Wei, W. Yu, R. Li, LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT, IEEE Trans. Ind. Inf. 16 (8) (2019) 5244–5253.

[7] I.V. Tetko, P. Karpov, R. Van Deursen, G. Godin, State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis, Nat. Commun. 11 (1) (2020) 1–11.

[8] L. Sun, C. Xia, W. Yin, T. Liang, P.S. Yu, L. He, Mixup-transformer: dynamic data augmentation for NLP Tasks, arXiv 1 (2019) 1–5, preprint arXiv: 2010.02394.

[9] K. Choi, J. Yi, C. Park, S. Yoon, Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines, IEEE Access, 2021.

[10] H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time-series forecasting: current status and future directions, Int. J. Forecast. 37 (1) (2021) 388–427.

[11] G. Petneházi, Recurrent neural networks for time-series forecasting, arXiv 1 (2019) 1–22, preprint arXiv: 1901.00069.

[12] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time-series through stochastic recurrent neural network, in: Proceedings of the 25th ACM *SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, July, pp. 2828–2837.

[13] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, T. Soderstrom, Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding, in: Proceedings of the 24th ACM *SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, July, pp. 387–395.

[14] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, N.V. Chawla, A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time-series data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, July, pp. 1409–1416. No. 01.

[15] B. Zong, Q. Song, M.R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding Gaussian mixture model for unsupervised anomaly detection, in: International Conference on Le*arning Representations*, 2018, February.

[16] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.K. Ng, MAD-GAN: multivariate anomaly detection for time-series data with generative adversarial networks, in: International Conference on Arti*ficial Neural Networks*, Springer, Cham, 2019, September, pp. 703–716.

[17] A.P. Mathur, N.O. Tippenhauer, SWaT: a water treatment testbed for research and training on ICS security, in: 2016 International Workshop *on Cyber-Physical Systems for Smart Water Networks (CySWater)*, IEEE, 2016, April, pp. 31–36.

[18] T. Nakamura, M. Imamura, R. Mercer, E. Keogh, Merlin: parameter-free discovery of arbitrary length anomalies in massive time-series archives, in: 2020 IEEE International Confere*nce on Data Mining (ICDM)*, IEEE, 2020, November, pp. 1190–1195.

[19] J. Audibert, P. Michiardi, F. Guyard, S. Marti, M.A. Zuluaga, Usad: unsupervised anomaly detection on multivariate time-series, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, August, pp. 3395–3404.

[20] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, Q. Zhang, Multivariate time-series anomaly detection via graph attention network, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, November, pp. 841–850. Multivariate time-series anomaly detection via graph attention network. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 841–850). IEEE.

[21] A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time-series, No. 5, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, May, pp. 4027–4035.

[22] Y. Zhang, Y. Chen, J. Wang, Z. Pan, Unsupervised deep anomaly detection for multi-sensor time-series signals, IEEE Trans. Knowl. Data Eng. 1 (2021), 1–1.

[23] M.G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, A. Walters, Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep, Sleep Med. 3 (2) (2002) 187–199.

[24] T. Tayeh, S. Aburakhia, R. Myers, A. Shami, An attention-based ConvLSTM autoencoder with dynamic thresholding for unsupervised anomaly detection in multivariate time-series, Machine Learn. Knowled. Extract. 4 (2) (2022) 350–370.

[25] I. Kovalenko, M. Saez, K. Barton, D. Tilbury, SMART: a system-level manufacturing and automation research testbed, Smart. Sustain. Manuf. Syst. 1 (1) (2017).

[26] S. Tuli, G. Casale, N.R. Jennings, Recurrent neural networks for time-series forecasting, arXiv 15 (2019) 1201–1214, preprint arXiv: 2201.07284.

[27] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50.