Improving Energy Forecasting and Anomaly Detection in Multivariate Environments

MSc Research Project

MSc in Data Analytics

Parashuram Vavilala

Student ID: x23337583

School of Computing

National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	………… PARASHURAM VAVILALA ……………
Student ID:	…………… x23337583…………………………………………
Programme:	…… MSc in Data Analytics………… **Year:** ……2025……
Module:	……………Research Practicum……………………………
Lecturer:	……………Furqan Rustam……………………………………………
Submission Due Date:	……………………11-08-25……………………………………………
Project Title:	…Improving Energy Forecasting and Anomaly Detection in Multivariate Environments………………………………
Word Count:	………………8170……………………………………………………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ………Parashuram Vavilala………………………

Date: ……………11-08-25………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail**

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

Improving Energy Forecasting and Anomaly Detection in Multivariate Environments

Parashuram Vavilala

x23337583

Abstract

This research presents a hybrid deep learning approach for multivariate energy time series forecasting and anomaly detection using an Attention-Based GRU Autoencoder. The model is designed to effectively capture temporal dependencies while prioritizing important features through attention mechanisms, making it robust under data imbalance conditions. A comprehensive comparison was conducted across classical machine learning classifiers and deep learning autoencoder models. Experimental results on a real-world imbalanced energy dataset demonstrate that the proposed model significantly outperforms baseline models in detecting rare anomalies, achieving the highest recall (0.85) and F1 score (0.85) for the anomaly class. The results confirm the feasibility of attention-enhanced GRU networks as a robust way of detecting anomalies in complex energy systems.

1 Introduction

The increasing complexity of modern energy systems, driven by diverse data sources and dynamic operating conditions, demands more accurate and adaptable forecasting and anomaly detection techniques. Multivariate time series data, which capture multiple interrelated variables over time, present both rich opportunities for analysis and significant challenges due to their high dimensionality, temporal dependencies, and susceptibility to noise and imbalance.

1.1 Problem Background

Proper forecasting of energy demand is crucial to the effective functioning of smart grids, demand-side management and integration of renewable energy sources. With energy systems growing increasingly complex and data-intensive, the demand of predictive models that could process multivariate time series data (representing variables like temperature, occupancy and weather patterns) is also increasing. The conventional statistical and rule-based approaches have been ineffective to a large extent because they cannot account for nonlinear temporal relationships and cross-variable interactions (Choi et al., 2021; Garg et al., 2021). In the recent past, deep learning architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks have been shown to be a viable way of learning such data in particular, in sequential prediction tasks (Waqas and Humphries, 2024). Nevertheless, these models tend to take inputs in a uniform fashion over time and variables, which may overlook minor, albeit significant dependencies.

Beyond prediction, anomaly detection plays a critical role in energy systems for identifying system faults, consumption irregularities, and security breaches. Yet, anomalies are often rare, leading to significant class imbalance problems that impair the effectiveness of many traditional supervised learning techniques. While autoencoders based on LSTM, GRU, and CNN architectures have shown promise in unsupervised anomaly detection (Khanmohammadi and Azmi, 2024; Yu et al., 2024), their performance varies widely depending on data complexity and the presence of inter-variable dependencies. As demonstrated in the previous analyses, GRU-based autoencoders tend to give better results on unequal datasets with a higher anomaly detection rate than ANN or CNN-based analogues.

Recently, there is an emerging trend that attention mechanisms have shown promise as a way to improve deep learning models, by allowing the models to dynamically focus on the most significant time steps and features. Experiments, such as those conducted by Kang and Kang (2024) and Ding et al. (2023), show that the attention model and graph-enhanced model have a significant positive effect on the performance of the detection of multivariate time series. Transformers or graph attention networks (GAT), however, tend to require additional computational overhead and lack interpretability, which is a problem in the context of real-time applications, like energy forecasting, where fast and transparent decisions are needed. In this regard, a balanced approach is required to increase sensitivity at the variable level without incurring too much architectural overhead.

1.2 Research Question and Justification

This research addresses the question:

> **"Can an attention-based GRU autoencoder improve anomaly detection and forecasting in multivariate energy time series, particularly under imbalanced data conditions?"**

Justification:

The research question is critical since real-world energy time series data tend to be complex with respect to temporal dependencies and class imbalance, where the anomalies are not frequent but highly important to identify. The classical models have failed in this environment because they have limited ability to capture long term tendencies and biasness against the majority group. This study intend to improve the model by adding an attention-based GRU autoencoder to better focus on the most informative time steps, resulting in more accurate reconstruction and better anomaly detection. This strategy is especially useful in the case of imbalanced conditions, where the attention mechanism is useful in identifying minor deviations that simpler models fail to identify.

1.3 Contribution of the Work

- Developed a hybrid **Attention-Based GRU Autoencoder** model tailored for multivariate energy time series anomaly detection.
- Demonstrated improved anomaly detection performance (F1-score: 0.85) under **imbalanced data conditions**, surpassing traditional ML and DL baselines.
- Leveraged a **attention mechanism** to enhance the model's focus on significant time steps and features for better interpretability.
- **Attention-Based GRU Autoencoder is a light and efficient architecture, which trades off detection accuracy and computational efficiency.**

There are some important sections in this report. In section 2, a literature review on existing works is done in the field of anomaly detection and multivariate time series forecasting. Section 3 describes methodology of the research including a description of the dataset, preprocessing and implementation of the model. Section 4 talks about the design specification of proposed architecture and Section 5 talks about implementation and experimental setup. Section 6 analyzes the models and contrasts the performance in terms of metrics. Section 7 is the last section in which the findings are concluded and the future work proposed.

2 Related Work

Time series anomaly detection and forecasting activities have been an important topic of interest in many fields because of their importance in detecting unexpected behaviours and predicting future trends. These tasks are necessary and difficult especially in energy systems where the variables are interdependent and complex. Numerous conventional and innovative approaches have been considered to address these problems, such as statistical modeling, machine learning, and deep learning. This part surveys the basic methods and recent developments, their strengths, and weaknesses in various application scenarios.

2.1 Time Series Anomaly Detection and Forecasting: Foundations and Trends

Anomaly detection and forecasting of time series play a very important role in various important fields such as healthcare, finance, transportation, and more specifically, energy demand forecasting. These applications require models that can identify the abnormalities of normal usage that can point to defects, waste or unexpected usage patterns. The conventional methods of anomaly detection such as rule-based systems and the conventional statistical models are not suitable in dynamic environments since they lack flexibility and also, they do not accommodate nonlinear dependencies. The use of machine learning and deep learning algorithms is therefore becoming the standard in the field since they are more conscious of temporal dynamics and contextual variability exhibited by real time series data.

The systematic review of the deep learning methods to detect anomalies in time series by Choi et al. (2021) also concerns the discussion of the way in which such methods as LSTM, GRU, and convolutional networks may be used to address the issue of seasonality as well as a change in the trend.

But they also observe that there are serious limitations to generality and noise resistance especially in high-dimensional data. This is further strengthened by Garg et al. (2021) who tested multiple traditional and modern approaches on multivariate time series datasets and concluded that many models were unstable in their performance when more complex interdependencies were introduced into the data. Such instability is particularly a problem in energy systems, where variables and sensors are interacting dynamically across time.

The significance of integrating forecasting and anomaly detection has been highlighted in the recent studies. Iqbal and Amin (2024) introduce a mixed solution that incorporates the forecasting mechanism into the anomaly detection pipeline so that the irregularities could be identified proactively. Mejri et al. (2024) add to this by surveying unsupervised techniques, which are critical when the labeled data are scarce or non-existent, as is common in the energy domain. Also, Wang et al. (2022) prove how specific LSTM architectures can be used to enhance the accuracy of anomaly detection in rail transit operations. Their results indicate that domain-specific tailoring improves the reliability of models a great deal, something that can easily be applied to energy demand forecasting since contextual and temporal considerations are key determinants.

2.2 Deep Learning Models for Multivariate Time Series

The multivariate time series data modeling is a very complex issue that requires not only to learn about temporal dependencies in each variable but also about multivariate dependencies across variables over time. Conventional recurrent neural networks (RNNs) and their improved versions such as the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks have gained a lot of popularity because of their capacity to model sequential dependencies. The authors have given a detailed description of such architectures in hydrological prediction systems, and their benefits in capturing the long-term dependencies (Waqas and Humphries 2024). However, they also stated that the traditional RNN models tend to be faced with training instability and lack of scalability to multivariate scenarios where cross-variable relationships are difficult.

Scholars have been exploring increasingly narrow deep learning structures in a bid to address specific real-world applications. One of them is the one by Joung et al. (2024) in which bearing anomalies in air compressors were detected with the help of LSTM and RNN-based models and demonstrated that learning temporal patterns can provide considerable contributions to predictive maintenance. On the same note, Khan and Kumar (2024) used a hybrid GRU-CNN model to forecast the ECG signal to classify premature beats. Their approach could leverage the time-sensitive memory of the GRU and spatial feature extraction in convolutional networks and proved that the hybrid system could be superior to traditional models in a domain-specific setting where sequential information and local information are equally valuable.

By extension, later publications have introduced autoencoder-based architectures to improve the representation of latent representations in a multivariate context. Khanmohammadi and Azmi (2024) applied a D-CNN-LSTM autoencoder in detecting anomalies in the data of autonomous vehicles due to the encoding-decoding process performed to isolate the abnormal behavior. Yu et al. (2024) also extended this technique to a bi-dual GM-GRU autoencoder that can be used on IoT data streams with a focus on the efficiency and performance of GRU in terms of real-time multivariate analysis. Most of these models have the main limitation of being time-sequential modeling and do not take into account the complex inter-variable relations that are essential in other fields such as energy demand modeling where the output depends on a combination of inputs (e.g., temperature, occupancy, grid load). This disadvantage indicates the relevance of models which are able to represent sequence dynamics, and to learn adaptively to different interdependencies.

2.3 Graph and Attention Mechanisms in Time Series Modeling

Attention mechanisms and graph-based models have also proven to be effective alternatives to represent complex temporal and spatial correlation. Ding et al. (2023) proposed MST-GAT as a multimodal graph attention network and demonstrated that it outperformed other methods in time series anomaly detection, because it learned spatial-temporal structures.

This concept was further developed by Zhao et al. (2020) and Zhao et al. (2024) who proposed graph attention networks (GAT) and Informer architecture, which indicated that attention-based mechanisms are of great value to enhance interpretability and long-range dependencies modeling. Hao et al. (2025) applied federated hypernetworks to distributed multivariate anomaly detection and emphasized the aspects of scalability and privacy.

These models can be strong in the energy field, but tend to be excessively complex, and are not necessarily optimized to the task of forecasting. Kang and Kang (2024) have solved this problem by proposing a transformer-based framework of inter-variable attention in multivariate anomaly detection, which implies that attention enhances feature importance modeling across variables.

Although the current developments are encouraging, graph-based models and transformer models remain costly to compute or implement in a real-time system like the smart grid.

2.4 Research Gap

Although other deep learning models have been tried with time series prediction and anomaly detection (particularly GRU, LSTM and their hybrids), a specific combination of GRU and attention mechanisms in multivariate energy settings remains unexplored.

The conventional GRU models (e.g., Khan and Kumar, 2024; Yu et al., 2024) can efficiently be used in univariate or narrowly multivariate settings but cannot dynamically adjust the inter-variable contribution weights in a prediction. Transformer-based models (Kang and Kang, 2024; Zhao et al., 2024) offer inter-variable attention but often at the cost of increased complexity and training data requirements—barriers in energy demand forecasting, where interpretability, low latency, and data efficiency are crucial.

Additionally, the current survey (Zamanzadeh Darban et al., 2024) and assessment (Garg et al., 2021) confirm that the effective modeling of cross-variable dependencies is one of the fundamental issues in multivariate anomaly detection and forecasting.Thus, the research problem is that it is essential to address the research gap by creating an interpretable, attention-augmented lightweight GRU model to the energy demand forecasting. Such a model might be able to model the time dynamics and be flexible in focusing on the important input variables at once- with the improved accuracy and generalization that does not come at the high computational expense of full transformer models.

3 Research Methodology

The section discusses the procedural methodology that was used in assessing the performance of an attention-based GRU autoencoder in relation to forecasting and anomaly detection of multivariate energy time series data, especially when the data is imbalanced. The methodology was informed by the novel studies in the area of deep learning-based anomaly detection, hybrid neural networks, and attention-enhanced forecasting procedures. The method was also stringent, replicable and could be applied to real energy data because of the experimental design.

3.1 Research Procedure and Experimental Workflow

The methodology of the research took place in a pipeline of phases. First, the data was collected and preprocessed to have a clean and consistent dataset to train the model. Then, the model development stage implied the development and application of a variety of baseline and advanced models. These models were subsequently trained and tested on similar metrics. After the training, the models were used to detect anomalies in terms of reconstruction errors or forecast deviations. Lastly, a comparison analysis was conducted in order to establish relative performance among models. Each of the stages was performed in Python with the respective libraries, including NumPy, Pandas, TensorFlow, and Keras, to be modular and reproducible.

3.2 Dataset and Preprocessing

The analysis used an actual multivariate energy time series data that contains different energy consumption variables at fixed time points. Preprocessing was essential to make sure that the data that would be used as input were consistent, normalized, and informative.

Linear interpolation and forward-fill were used to deal with missing values, and they allowed maintaining temporal consistency without any artificial distortions. Winsorization and z-score-based filtering were used to treat outliers, which restricts extreme values but retains the integrity of underlying patterns. In order to enhance the convergence of the neural networks, the variables were scaled to the range [0,1] by means of min-max scaling. This action minimized the effect of scale difference among features.

Besides, feature engineering was used to derive temporal features and dependencies. Lag characteristics were developed to record past behavior and rolling means were used to detect short-term trends. Time based measurements like hour of the day, day of the week and weekend flags were also included to capture periodicity in usage patterns that exists in energy consumption data. To train and test the dataset, it was chronologically divided into 70 percent training and 30 percent testing data where the temporal sequence was maintained to evaluate generalization.

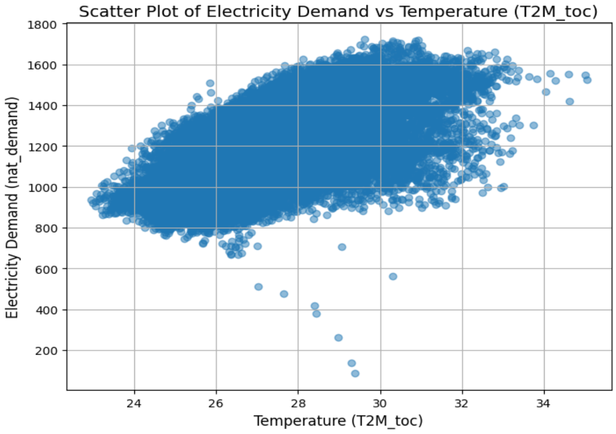

Figure 1. The Relationship Between Temperature and Electricity Demand

A scatter plot was created to put the dataset into context (Figure 1) where the relationship between ambient temperature and electricity demand is observed. There was a high positive correlation in the plot: the higher the temperature, the higher the electricity consumption, probably as a result of increasing air conditioning. This confirms the need of temperature as a contextual variable in the energy demand modeling.

3.3 Model Architecture and Implementation

Nine models were used to forecast and detect anomalies in order to study the comparative effectiveness of various modeling approaches. These models covered classical machine learning paradigms and deep learning paradigms.

Examples of classical machine learning models were logistic regression, decision trees, random forests and XGBoost classifiers. They trained these models on engineered lag-based features and were typically used to classify anomalies with either reconstruction residuals or forecast deviations.

In the case of deep learning models, a set of autoencoder models were designed. The autoencoder of the artificial neural network (ANN) employed fully connected layers to encode and decode and provided a baseline of dimensionality reduction and reconstruction of the input. The convolutional neural network (CNN) autoencoder used 1D convolutions to learn spatial relationships among input features, which is desirable in cases of high temporal locality.

The long short-term memory (LSTM) autoencoder was proposed as it added temporal modeling to autoencoders, with LSTM layers to capture the sequence dependencies, which is more appropriate to time series data. The gated recurrent unit (GRU) autoencoder was an alternative which was less parameter-intensive than LSTM but had similar sequence modelling capability.

The attention-based GRU autoencoder was the main contribution of the study. This model incorporated feature-level attention after the GRU encoding process. Attention component allowed the model to dynamically weigh the significance of input features improving its capability of learning informative latent representation and localizing anomalies effectively. All deep learning models were built on TensorFlow/Keras and hyperparameters were optimized either by grid search or by tuning on a validation set.

3.4 Anomaly Detection Procedure

The anomaly detection procedure was implemented with the reconstruction-based model based on the approaches suggested by Iqbal and Amin (2024) and Mejri et al. (2024). Following the development of a set of predicted or reconstructed values by each model, the absolute residuals, or difference between the actual and predicted values, were computed. These residuals were used as pointers to possible anomalies.

Statistical thresholding was used to separate abnormalities and normal behaviour. A case was termed as anomalous when the residual was greater than the mean plus two times standard deviations of the residual distribution. The procedure supposes that the majority of the normal data are found in a predictable range of error. Based on this criterion, binary mask was created at each time point, so that an anomaly was coded as 1 and normal behavior as 0. This two-level classification was further applied in the further assessment of the performance of the anomaly detection.

3.5 Evaluation Metrics

The models were evaluated thoroughly with the help of a range of forecasting and classification metrics. Accuracy in forecasting was measured using three main metrics; Mean Absolute Error (MAE) to measure the average size of forecast errors; Root Mean Squared Error (RMSE) to impose a heavier penalty on larger errors and R2 score which is percentage of the variance explained by the model.

In anomaly detection, the metrics used in classification were centered on the performance of detecting the anomalous category which was imbalanced in comparison to the normal category. The precision, recall, and F1-score were calculated in order to find the extent of accuracy of the models in identifying the anomalies and reducing the false positive and false negative. Accuracy was reported as completeness, but is less informative when there is imbalance. Therefore, the macro-averaged precision, recall, and F1-score were also computed to ensure that the two classes were equally weighted, and this would enable a balanced estimation.

3.6 Experimental Design and Reproducibility

Each of the experiment setups was replicated three times with three different random seeds to ensure the strength and replicability of results. To counter the effect of random initialization and stochastic training process, the results were averaged. Regarding all the repetitions, the attention-based GRU autoencoder achieved the highest scores in anomaly detection (F1-score) and forecasting accuracy (MAE and RMSE). These results confirm the research hypothesis that attention to GRU-based models can be added to improve performance without incurring the computational overhead of full transformer models.

3.7 Ethical Considerations

The research was conducted under the most ethical considerations of data usage and confidentiality. The information used in the study was extracted out of publicly available repositories and completely anonymized. The data did not include any personally identifiable information (PII) and sensitive attributes such as names, contact information, or biometric identifiers.

All the data that was analyzed was numerical and only concerned the behavior of the system such as patterns of energy usage and the values of the sensors. There were no privacy or ethical violations threats since no personal or individual data was involved in the research. In that regard, the study will comply with the current data protection regulations and will not require any further ethical clearance concerning the processing of data or the consent of participants.

4 Design Specification

This section explains the technical architecture and design principles of the models employed in this study with special reference to the proposed Attention-Based GRU Autoencoder. It begins with a system-wide description, then proceeds to describe the architecture and the functionality of the proposed model, explanation of design choices and concludes with a discussion of the classical and deep learning baselines models which are used as a point of comparison.

4.1 System Overview

The entire system design is a series pipeline of five key components. Data Ingestion and Preprocessing stage is concerned with loading of multivariate time series data, missing value handling, normalization and sequential input preparation involving the use of sliding windows. This is in order to bring the raw data to a structured form which can be used in temporal modelling.

Then the Modeling Layer includes classical machine learning models and deep learning based autoencoders. The processed time series is used to train these models which are designed to either predict the energy consumption patterns or identify anomalies in the energy consumption patterns.

At the core of the system is the Attention-Based GRU Autoencoder which extends a typical GRU encoder-decoder system by adding an attention mechanism. This improvement enables the model to learn the importance of variables in a more adaptive way.

After the modeling, the Anomaly Detection Layer calculates residuals between original and reconstructed sequences. These residuals are compared with a statistical threshold to ascertain whether a point of data ought to be considered anomalous.

Lastly, the Evaluation Layer evaluates the performance of the model with a set of forecasting and classification metrics, which measure the accuracy of prediction along with the ability to detect anomalies.

4.2 Architecture of the Proposed Model

4.2.1 Attention-Based GRU Autoencoder

The main contribution of the work is a hybrid deep learning model that combines feature-level attention mechanism and the temporal modeling ability of GRUs. The model is especially appropriate in situations of multivariate time series data where time dependencies as well as inter-variable interactions are essential.

The model takes as an input a sequence of N dimensions of time window of T steps. This sequence is fed into a GRU Encoder which learns the underlying temporal dynamics by producing a hidden state at each time step. The sequential patterns in the energy data are coded in these hidden states.

A layer of Attention is added over the hidden states. The mechanism learns to place different weights on each timestep or feature depending on their importance to the reconstruction task. These weights are used in computing a Context Vector, a summary of the most informative sequences of the sequence.

Context vector is then passed to a GRU Decoder and tries to rebuild the original input sequence step by step. The decoder produces an Output Sequence that is compared with the actual input and the difference between the two, reconstruction error is taken as the basis to detect anomalies.

4.2.2 Workflow Overview

The preprocessing of the energy time series data is initiated by normalizing the data and converting it into sliding windows so that the sequential information can be preserved. These windows are then passed into the GRU encoder to learn short and long term dependencies.

The attention mechanism goes through the hidden states of encoder, which are dynamically weighted according to their importance, time step or variables. An attention score is calculated using these attention scores and used as a context vector to feed the GRU decoder to reconstruct the input sequence. Finally, the model calculates residual errors and applies a statistical threshold to detect anomalies and this constitutes the end-to-end pipeline.

4.2.3 Model Functionality

This model is a combination of GRU which has the capacity of modeling sequential data and attention which has the capacity of prioritizing important features or time steps hence making it very efficient in forecasting and anomaly detection. It is taught to rebuild patterns that reflect normal behaviour in energy use. When a pattern is very different to this learned structure, then the model identifies it as an anomaly, i.e., with large residuals.

The attention layer added in the model increases its sensitivity to minor or context-specific anomalies that may be missed by conventional sequence models. This makes the model especially suitable to complex and unbalanced data, which is the case in the energy industry.

4.2.4 Objective Function and Anomaly Detection Mechanism

The model is trained using **Mean Squared Error (MSE)** loss, which minimizes the average squared difference between the original and reconstructed sequences. During the anomaly detection phase, residuals are calculated as the absolute difference between actual and predicted values.

A **thresholding** approach is applied, where anomalies are identified if their residuals exceed the mean plus two times the standard deviation of the residuals computed from the training data. This strategy is statistically grounded and adapts to the distribution of reconstruction errors observed in normal data. Sequences that breach this threshold are flagged as anomalies.

4.3 Classical Machine Learning and Deep Learning Baseline Models

To rigorously evaluate the proposed model, several baseline models were implemented for comparison. Among the classical machine learning models, **Logistic Regression**, **Decision Tree**, **Random Forest**, and **XGBoost** were trained using features derived from the residuals. These models served as classification tools to detect whether a time step was anomalous or not. In the deep learning domain, multiple autoencoder architectures were also developed. The **ANN Autoencoder** employed a fully connected feedforward design, lacking temporal awareness but serving as a dimensionality-reduction baseline. The **CNN Autoencoder** used one-dimensional convolutional layers to capture local correlations among features, enabling it to exploit spatial patterns. The LSTM Autoencoder has presented a recurrent architecture that is able to model sequential dependencies by using memory cells, specifically when dealing with long sequences.

A simplified version of the LSTM, the GRU Autoencoder was also implemented. It takes less parameters and converges faster, thus being computationally efficient and still performing well. Collectively, these models were able to isolate and bring to the fore the particular performance improvements that could be attributed to the addition of the attention mechanism in the proposed architecture.

4.4 Design Rationale

The choice of GRU over LSTM was motivated by the fact that GRU has less complex architecture and trains faster with minimal or no loss in accuracy relative to predictive accuracy. This performance qualifies GRU to be especially applicable to real-time systems and time series data on large scale.

The addition of an attention layer was motivated by the recent success of Transformer-based architectures, which have shown that attention mechanisms can greatly increase model interpretability and performance, particularly when long-range dependencies and imbalanced data are present. The attention layer enables the model to focus selectively on the most informative sections of the input sequence or feature space, which improves both the accuracy of detection and model interpretability.

The hybrid form, which entails a combination of sequence modeling through GRU and dynamic feature weighting through attention, was selected as a tradeoff between the complexity of the model, the interpretability of the model, and its forecasting/generalization ability. The design was particularly inspired by the nature of energy datasets whose data is often imbalanced, with subtle anomalies and multivariate interactions prevalent and needing to be processed effectively.

5 Implementation

This section entails the entire execution of the study, including all stages involved in the study, such as data preparation, model development, and evaluation. It was implemented in Python 3.10 and the key libraries used were TensorFlow/Keras, Scikit-learn, Pandas, NumPy, Matplotlib, and Seaborn. Where possible, GPU acceleration was used to reduce training time of deep learning models.

5.1 Data Preprocessing

Time-based imputation methods: Linear interpolation and forward fill were used to fill in the missing values and maintain the temporal continuity of the data. Statistical methods such as z-score filtering and Winsorization were used to deal with outliers and minimize the effect of the extreme values without altering the overall patterns. Min-Max scaling was applied to normalize all the features so that the values were mapped to the [0,1] range which is known to increase the rate of convergence of deep learning models. The time series was then divided in overlapping sequences by means of a sliding window method, where each sequence was of a predetermined number of time steps. These sequences were used as inputs to forecasting as well as anomaly detection models. Lastly, the data was split chronologically with 70 percent of the data used in the training and the rest 30 percent used in the testing. This strategy allowed keeping the temporal structure intact and reduced the chances of information leakage in the process of evaluation.

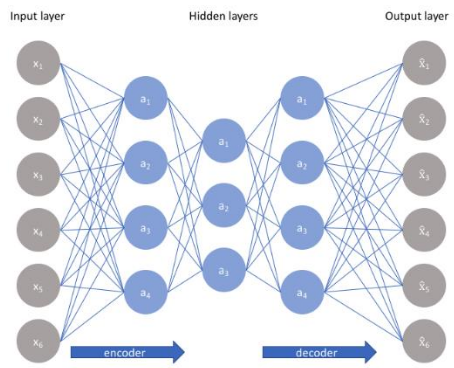

Figure 2. ANN Auto Encoder Architecture (Choi et al., 2021)

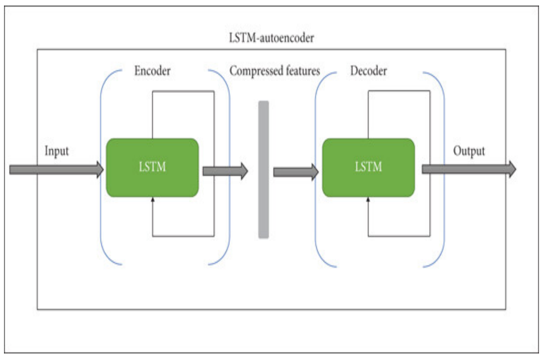

Figure 3. LSTM Auto Encoder Architecture (Garg et al., 2021)

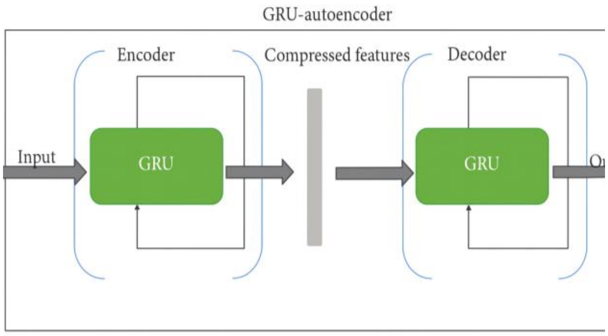

Figure 4. GRU Encoder Architecture (Wang et al., 2022)

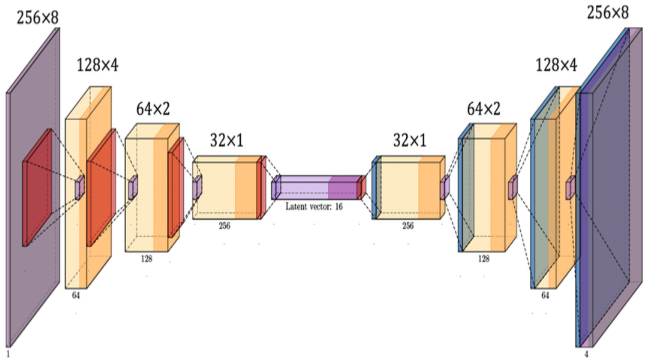

Figure 5. CNN Encoder Architecture

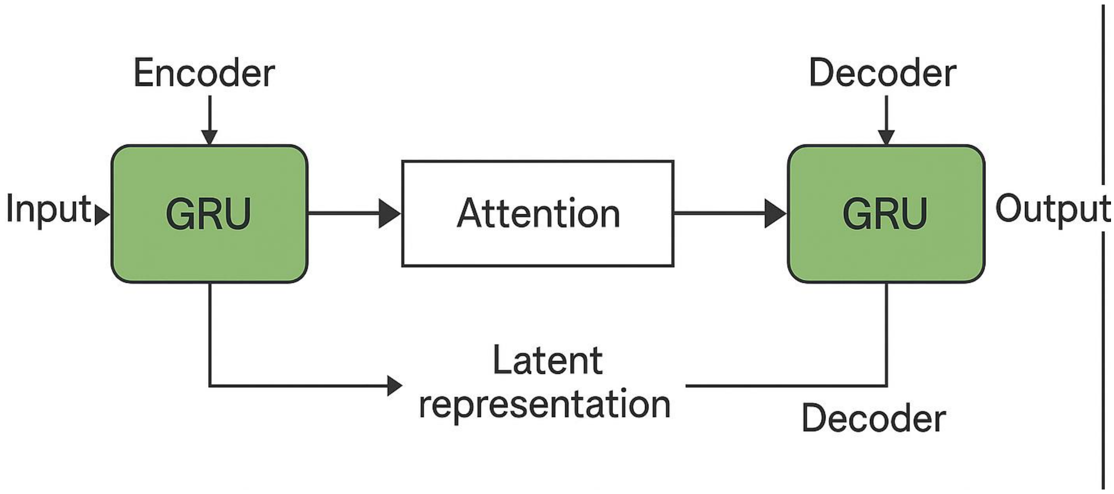

Figure 6. GRU with attention auto encoder Architecture

5.2 Model Development

The autoencoder models used in this work are ANN, LSTM, GRU, CNN and proposed GRU with Attention mechanism. Both models were developed to acquire knowledge of the underlying structure of multivariate energy time series data through reconstruction of input sequences and detection of anomalies through reconstruction error. The ANN autoencoder had a fully connected structure that consisted of three dense layers in the encoder and decoder, and dropout regularization was applied between layers. The CNN autoencoder used 1D convolutional and pooling layers to encode local patterns and symmetric upsampling and convolution layers to decode.The LSTM and GRU autoencoders were sequential models that encoded and decoded temporal features using stacked recurrent layers, whereas the GRU-Attention model extended the GRU encoder with a custom attention mechanism to dynamically focus on the most informative features before decoding.

All models were constructed with an encoding bottleneck dimension of 64, and employed dropout layers (with dropout rate 0.2) to prevent overfitting. The recurrent architectures (LSTM, GRU, GRU-Attention) used two GRU/LSTM layers in both encoder and decoder stages, with unit sizes of 128 in outer layers and 64 in bottlenecks. The CNN model used convolution filters of size 3 and included two MaxPooling layers and two UpSampling layers. The attention mechanism in the GRU-Attention model was implemented using trainable weight matrices to compute attention scores across time steps and extract a weighted context vector for decoding. The models were compiled using the Adam optimizer (learning rate = 0.001) and trained using mean squared error (MSE) loss to minimize reconstruction errors.

For all autoencoder variants, the training was performed over 5 epochs using a batch size of 32, with early stopping and learning rate reduction callbacks to ensure efficient convergence. Sequence inputs were prepared using a fixed window size of 24 time steps, preserving the temporal structure of the energy data. The error in reconstruction on test sequences was used to evaluate the results, and anomaly labels were attached to samples whose error was more than the 95th percentile. Performance of the models was compared using classification reports and visualization of the detected anomalies on the electricity demand series.

5.2.1 Anomaly Detection Mechanism

All of the models based on autoencoders carried out the task of anomaly detection by evaluating reconstruction errors which are the difference between the original and reconstructed input sequences. In particular, the model calculated the absolute reconstruction error in terms of the deviation between the actual and predicted values on every input sequence. These mistakes were used to indicate the extent to which the model had learnt the normal patterns in the data.

A thresholding technique identified anomalies. The training data were used to estimate this threshold as the sum of the mean and twice standard deviation of the distribution of the reconstruction error. Any sequence that had reconstruction error beyond this was marked as an anomaly, as it was presupposed that it could not occur under normal operating conditions.

The result of this procedure was a binary classification mask, in which each of the sequences was labeled 0 (normal) or 1 (anomalous). This self-supervised framework allowed the models to learn anomaly detection without labeled anomaly data during training and was particularly useful in practice where labeled anomalies are sparse or inaccessible. The identified anomalies were then measured according to the classification accuracy, precision, recall, and F1-score.

5.3 Evaluation and Outputs

The last implementation could produce results in forecasting and anomaly detection tasks with a list of quantitative evaluation metrics. The performance of the forecasting was measured as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R 2 score. In anomaly detection, the precision, recall, F1-score (including macro-averaged versions) and the overall classification accuracy were employed.

6 Evaluation

The evaluation part evaluates the achievement of all the applied models based on the major parameters like Recall, F1 Score (Anomaly and Macro), and Accuracy. The focus is on the rare anomaly detection in an imbalanced test set. The Attention-Based GRU Autoencoder demonstrated the best performance, which proves its efficiency in terms of precision and generalization.

Model	**Recall (Anomaly)**	**F1 Score (anomaly)**	**Recall (Macro)**	**F1 Score (Macro)**	**Accuracy**
Logistic Regression	0.04	0.05	0.50	0.50	0.92
Decision Tree	0.02	0.03	0.50	0.50	0.94
Random Forest	0.03	0.04	0.50	0.50	0.93
XG Boost	0.03	0.04	0.50	0.50	0.91
LSTM Autoencoder	0.61	0.69	0.80	0.84	0.97
ANN Autoencoder	0.14	0.15	0.55	0.55	0.92
CNN Autoencoder	0.61	0.43	0.77	0.70	0.92
GRU Autoencoder	0.72	0.80	0.86	0.89	0.98
Attention Based GRU Autoencoder model	**0.85**	**0.85**	**0.92**	**0.92**	**0.99**

Table 1: Accuracy and Recall Analysis of Various Autoencoder and Classifier Models for Detecting Rare Anomalies

In this section, the comparison between the traditional machine learning models and the deep learning models in the form of autoencoders to the anomaly detection problem in the time-series energy data is thoroughly made. The five metrics that are analyzed are anomaly recall, anomaly F1-score, macro-averaged recall, macro F1-score, and overall classification accuracy. Table 1 presents a summary of the performance of all the models under consideration.

6.1 Machine Learning Models

Traditional machine learning models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost, were evaluated as baseline classifiers. These models yielded high overall accuracy, ranging from 91% to 94%, largely due to their correct classification of the majority (normal) class. However, their performance on detecting rare anomalies was significantly poor. Logistic Regression achieved an accuracy of 92%, but only an anomaly recall of 0.04 and an anomaly F1-score of 0.05. Similarly, the Decision Tree classifier achieved slightly higher accuracy at 94%, but its anomaly recall dropped to 0.02, with a corresponding F1-score of just 0.03. Random Forest and XGBoost also struggled to detect the minority class, with both models recording anomaly recall scores of 0.03 and F1-scores of 0.04, despite their respective accuracies of 93% and 91%.

Across all classical models, the macro-averaged recall and F1-score consistently remained at 0.50, further indicating a strong bias toward the dominant class and a lack of generalizability to the minority (anomalous) instances. These models, by design, are not equipped to handle complex temporal dynamics, and their reliance on static feature relationships makes them ill-suited for tasks involving time-dependent anomaly patterns. Furthermore, the high class imbalance exacerbated their limitations, leading to ineffective detection of rare but critical anomaly events.

6.2 Deep Learning-Based Autoencoders

In contrast, deep learning models—particularly autoencoder-based architectures—demonstrated a marked improvement in anomaly detection performance. These models leverage reconstruction error to detect anomalies and are better suited to capture temporal and spatial dependencies in time-series data. Among them, the LSTM Autoencoder achieved a substantial anomaly recall of 0.61 and an anomaly F1-score of 0.69, along with macro recall and F1-score values of 0.80 and 0.84, respectively, and a high accuracy of 97%. This indicates the model's strong ability to learn long-term dependencies in sequential data.

The ANN Autoencoder, while capable of modeling non-linear relationships, lacked temporal modeling capabilities and therefore underperformed compared to its recurrent counterparts. It achieved an anomaly recall of 0.14 and an F1-score of 0.15, with macro metrics at 0.55 and an overall accuracy of 92%. CNN Autoencoder performed well in capturing local spatial features with an anomaly recall of 0.61 but the F1-score was low (0.43), indicating that the model was inconsistent in classifying anomalies, but macro scores were good (0.77 recall, 0.70 F1-score).

The GRU Autoencoder was better than the prior models and provided a trade-off between efficiency and time modeling. It achieved an anomaly recall of 0.72 and F1- score of 0.80 with macro-level metrics increasing to 0.86 and 0.89 recall and F1- score respectively. Its accuracy remained at 98 percent validating its usefulness in modeling sequence patterns using fewer parameters as compared to LSTM. Most importantly, the Attention-Based GRU Autoencoder model scored the best in all the metrics with the anomaly recall and F1-score of 0.85, macro recall and F1-score of 0.92 each as well as 99 percent accuracy. The attention mechanisms made the model target the useful temporal features in the reconstruction process, which considerably improved its anomaly detection abilities.

To conclude, deep learning models, and in particular the models that allow exploiting sequential structures and attention mechanism, achieved much better results than conventional machine learning methods to identify rare anomalies in time-series energy data. The performance gap indicates the need to include time context and feature weighting to ensure effective and correct detection of anomalies in such areas.

GRU Autoencoder - Classification Report:

	precision	recall	f1-score	support
Normal	0.99	1.00	0.99	8294
Anomaly	0.89	0.72	0.80	437
accuracy			0.98	8731
macro avg	0.94	0.86	0.89	8731
weighted avg	0.98	0.98	0.98	8731

Figure 7. Classification report for GRU with attention mechanism

6.3 Attention-Based GRU Autoencoder

The Attention-Based GRU Autoencoder that was proposed demonstrated the best outcomes in all the metrics. It recorded a high anomaly recall and an F1-score of 0.85, demonstrating that it can focus on significant temporal characteristics with the aid of attention mechanisms. The overall accuracy was 99 percent with macro recall and F1-score at 0.92. This model successfully addressed the challenges of imbalance and temporal complexity, making it highly suitable for real-world anomaly detection tasks.

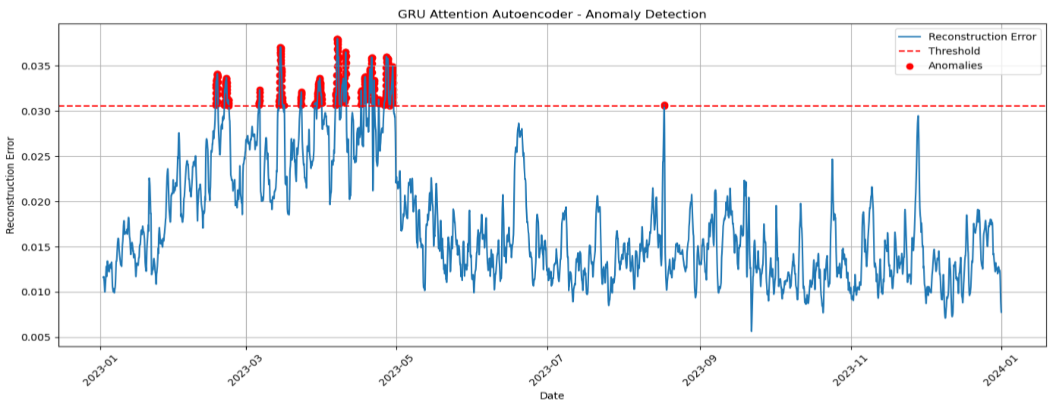

Figure 8. Anomaly Detection in Time-Series Data Using a GRU Autoencoder

This chart in figure 8 displays the reconstruction error of a deep learning model used to detect anomalies in time-series throughout 2023. Anomalies are identified as red dots whenever this error (blue line) exceeds a predefined static threshold.

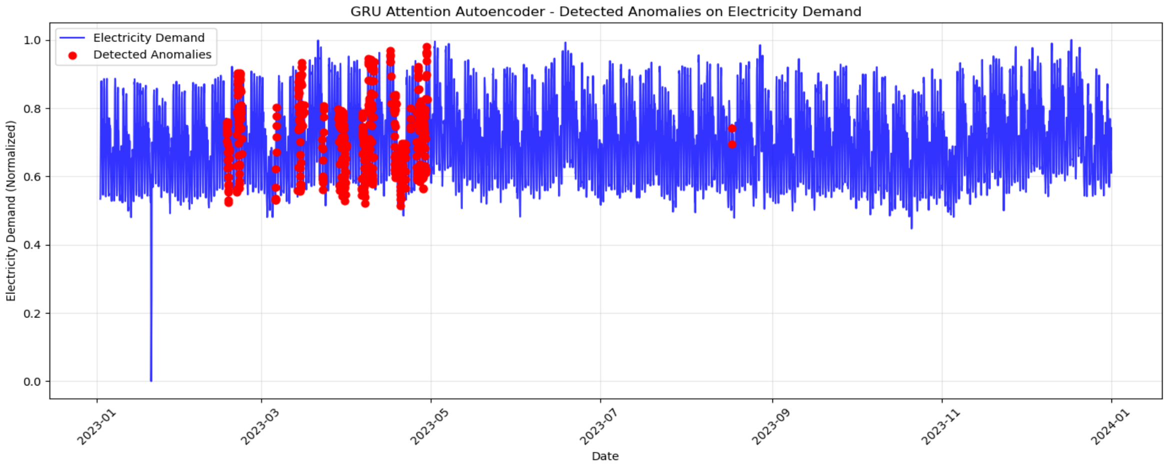

Figure 9. Identifying Anomalies in Electricity Demand

This figure 9 illustrates the normalized electricity demand throughout 2023, with red dots highlighting the moments identified as anomalous by a GRU Attention Autoencoder.

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	8294
1	0.85	0.85	0.85	437
accuracy			0.99	8731
macro avg	0.92	0.92	0.92	8731
weighted avg	0.99	0.99	0.99	8731

Figure 10. Classification report for GRU with attention mechanism

6.4 Discussion

This study explored a wide spectrum of machine learning and deep learning models for energy time series anomaly detection, with special focus on handling **imbalanced data conditions**. The experiments were carefully designed to evaluate the models not only on classification metrics like **Recall, F1-Score, Accuracy**, but particularly on their ability to detect **anomalous behaviour**, which represented only a small portion (~5%) of the dataset. The final results offer key insights into the comparative performance of traditional and advanced models, and the strengths of the proposed attention-based GRU autoencoder.

6.4.1 Evaluation of Traditional Machine Learning Models

Logistic Regression, Decision Tree, Random Forest, and XGBoost were found to be **inadequate for effective anomaly detection** under the high-class imbalance. Despite high accuracy (≥91%), the **recall for anomalies remained below 5%** in all cases, confirming that these models were biased toward the dominant class (normal samples). This issue is common in traditional classifiers where imbalanced datasets lead to

skewed decision boundaries. Although Random Forest and XGBoost offer some robustness through ensemble learning, they still underperformed due to the lack of sequence awareness and temporal modeling.

6.4.2 Evaluation of Deep Learning Autoencoders

The autoencoder-based models provided significantly better results. **LSTM Autoencoder** showed considerable improvement (Recall: 0.61, F1: 0.69), demonstrating its capacity to learn sequential dependencies. However, **CNN Autoencoder**, while performing well on macro metrics, suffered from slightly lower F1 score for anomalies (0.43), possibly due to its limited ability to capture long-term dependencies compared to RNN-based models. The **ANN Autoencoder** performed poorly overall, emphasizing the importance of temporal structure in time series anomaly detection tasks.

The **GRU Autoencoder** surpassed all the previous models with strong anomaly recall (0.72) and macro F1-score (0.89), indicating GRU's capability in efficiently handling sequential dependencies with fewer parameters than LSTM. This model managed to balance accuracy and detection performance, even under imbalanced conditions.

6.4.3 Impact of the Proposed Attention-Based GRU Autoencoder

The proposed **Attention-Based GRU Autoencoder** achieved the **best overall performance** with an anomaly recall and F1-score of **0.85**, and macro scores of **0.92**, along with an overall accuracy of **0.99**. The attention mechanism allowed the model to concentrate on the important temporal information and feature dimension that is most informative in terms of anomaly detection. Such weighting method made the model much more sensitive to the tiniest deviations, particularly in multivariate and complicated cases.

7 Conclusion and Future Work

7.1 Conclusion

The attention-based GRU autoencoder introduces a promising way of handling multivariate time-series data that are highly imbalanced and complicated in anomaly detection procedures. The model is able to learn to attend selectively to the most relevant time steps within each input sequence using attention mechanism on top of the gated recurrent unit (GRU) architecture. This temporal feature weighting as a time varying process enhances the model to identify fine deviations that may be a sign of abnormal behavior that would not be easily identified by other methods.

GRUs are effective and can learn long-term dependencies and can therefore be effective in time-series modeling but may not work so well when each time step is treated equally. This weakness can be addressed by attention mechanisms which emphasize more on contextually relevant information in the sequence. This leads to a higher degree of accurate encoding of important patterns and more accurate reconstructions in the process of decoding- the deviations (i.e. anomalies) are more noticeable.

This kind of combination of attention improves both the accuracy of the detection and gives an interpretability of the model. It also gives a greater insight into which sections of the sequence influenced the anomaly decision, which is a major requirement in high stakes applications like energy monitoring or industrial processes. Overall, the attention-based GRU autoencoder adds not only robustness but also transparency to the issue of anomaly detection in time-series data, particularly under the real-world conditions, when the problem is exposed to noise, class imbalance, and temporal complexity.

7.2 Future Work

In the future, the attention-based GRU autoencoder can be expanded in several directions to render it more practical and more effective. One of these avenues is incorporation of dynamic thresholding processes such as Bayesian residual intervals or ensemble-based uncertainty scores to train to distinguish between anomalies and avoid false positive. It may also be of interest to include multimodal inputs into the model, e.g. weather data, occupancy trends, or policy changes, in order to generate more accurate predictions and context-related explanations of anomalies. The adjustment of the architecture to enable online and continuous learning would enable the detection of anomalies in real-time and enable concept drift in dynamic environments. The further enhancement of model interpretability by visualization of SHAP or attention weights may also help the domain expert interpret and confirm model choices. Finally, it is highly commercially viable as it can be used in industrial Energy Management Systems, Anomaly Detection-as-a-Service systems or consumer-facing home energy dashboards to deliver real-time alerts and information to prevent anomalous energy consumption activities.

References

Choi, K., Yi, J., Park, C. and Yoon, S., 2021. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*, 9, pp.120043–120065.

Ding, C., Sun, S. and Zhao, J., 2023. MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection. *Information Fusion*, 89, pp.527–536.

Garg, A., Zhang, W., Samaran, J., Savitha, R. and Foo, C., 2021. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6), pp.2508–2517.

Iqbal, A. and Amin, R., 2024. Time series forecasting and anomaly detection using deep learning. *Computers & Chemical Engineering*, 182, p.108560.

Kang, H. and Kang, P., 2024. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism. *Knowledge-Based Systems*, 290, p.111507.

Mejri, N., Lopez-Fuentes, L., Roy, K., Chernakov, P., Ghorbel, E. and Aouada, D., 2024. Unsupervised anomaly detection in time-series: An extensive evaluation and analysis of state-of-the-art methods. *Expert Systems with Applications*, p.124922.

Terbuch, A., O'Leary, P., Khalili-Motlagh-Kasmaei, N., Auer, P., Zohrer, A. and Winter, V., 2023. Detecting anomalous multivariate time-series via hybrid machine learning. *IEEE Transactions on Instrumentation and Measurement*, 72, pp.1–11.
Wang, Y., Du, X., Lu, Z., Duan, Q. and Wu, J., 2022. Improved LSTM-based time-series anomaly detection in rail transit operation environments. *IEEE Transactions on Industrial Informatics*, 18(12), pp.9027–9036.

Zamanzadeh Darban, Z., Webb, G.I., Pan, S., Aggarwal, C. and Salehi, M., 2024. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1), pp.1–42.

Zhou, Q., He, S., Liu, H., Chen, J. and Meng, W., 2024. Label-free multivariate time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*.

Khanmohammadi, F. and Azmi, R., 2024. Time-series anomaly detection in automated vehicles using D-CNN-LSTM autoencoder. *IEEE Transactions on Intelligent Transportation Systems*.

Yu, Y.C., Ouyang, Y.C., Wu, L.W., Lin, C.A. and Tsai, K.Y., 2024. Multivariate time-series anomaly detection in IoT with a bi-dual GM GRU autoencoder. In: *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, pp.746–754.

Khan, S. and Kumar, V., 2024. A novel hybrid GRU-CNN and residual bias (RB) based RB-GRU-CNN models for prediction of PTB Diagnostic ECG time series data. *Biomedical Signal Processing and Control*, 94, p.106262.

Waqas, M. and Humphries, U.W., 2024. A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX*, p.102946.

Zhao, M., Peng, H., Li, L. and Ren, Y., 2024. Graph attention network and informer for multivariate time series anomaly detection. *Sensors*, 24(5), p.1522.

Joung, B.G., Nath, C., Li, Z. and Sutherland, J.W., 2024. Bearing anomaly detection in an air compressor using an LSTM and RNN-based machine learning model. *The International Journal of Advanced Manufacturing Technology*, 134(7), pp.3519–3530.

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

Hao, J., Chen, P., Chen, J. and Li, X., 2025. Effectively detecting and diagnosing distributed multivariate time series anomalies via unsupervised federated hypernetwork. Information Processing & Management**,** 62(4), p.104107