Anomaly Detection in Time Series: A Comparative Analysis of Machine Learning Techniques

***Abstract*—Anomaly detection in time series data is vital for maintaining system integrity across various domains. This paper compares the collected models of machine learning for anomaly detection namely Isolation Forest, OneClass SVM and Local Outlier Factor. These models were evaluated for their effectiveness in identifying anomalies in a power factor dataset. The collected results reveal the advantages and disadvantages of each model and emphasize that choosing an appropriate method depends on the characteristics of the dataset and operational task. The study contributes to the understanding of how to enhance the efficiency of anomaly detection in time series data.**

***Index Terms*—Anomaly Detection, Time Series, Machine Learning, Isolation Forest, OneClass SVM, Local Outlier Factor**

I. Introduction

An important class of methods in this respect is used widely within the fields in which cybersecurity, health monitoring, and industry management, among other things, strive to detect patterns that are considered on the measure of anomalous deviation. These anomalies might be errors, unusual events, or significant threats like fraud or network intrusions. The fast and accurate detection of anomalies allows organizations to be proactive toward potential catastrophic events and, hence, to maintain system integrity and operational continuity.

Time series data generally comprise a series of values over a single or multiple periods. The values or events are recorded at equal intervals of time. Time series data analysis aims to recognize the underlying patterns or changes, trends, and seasonal variations that are present. Indeed, fields such as economics, meteorology, and engineering depend so much on time series analysis to predict the future from the given historical data. As a result of non-stationary, high-frequency, and noise characteristics, data complexity subsequently brings a problem both in analysis and forecasting.

On the one hand, anomaly detection, when combined with time series analysis, improves the ability to look at time-dependent data for irregularities that otherwise might not be as obvious. Such a combination is essential, especially for dynamic environments in which data constantly changes and the cost of a delay in detecting an error may be high. Advanced algorithms, such as neural networks or machine learning models, can be applied over time series data for analysts to detect exceptional patterns, predict possible interruptions, and take timely corrective actions.

The primary contribution of this paper is the comprehensive evaluation of anomaly detection techniques applied to time series data. Specifically, three machine learning models—Isolation Forest, One-Class SVM, and Local Outlier Factor—were trained on both univariate and multivariate time series data. Additionally, an autoencoder with CNN layers was employed to detect anomalies in both univariate and multivariate contexts. By comparing the number of anomalies detected across these models, the study provides insights into the effectiveness of each method in handling different types of time series data, thereby advancing the understanding of model performance in diverse data scenarios.

This paper aims to develop a methodology for evaluating anomaly detection in time series data. The significance of this work lies in its thorough approach, comparing the number of anomalies detected using various machine learning and deep learning models on the Current Voltage dataset. The structure of the paper is as follows: Section 2 offers an in-depth literature review of relevant studies, focusing on anomaly detection using different model approaches. Section 3 describes the proposed system. Section 4 presents the findings of the experimental evaluation, and Section 5 provides concluding remarks and discussions.

II. Related Work

A. Machine learning

This study [1] shows the Microsoft dataset used for evaluating the SR-CNN model achieved an F1-score of 0.537. This highlights the challenges in anomaly detection, as manual labeling is impractical and the constantly changing nature of data complicates the process. Consequently, supervised models often prove insufficient in effectively detecting anomalies under these conditions.

In a comprehensive evaluation of anomaly detection techniques for time series data [2], LSTM-AD recorded an accuracy of 0.999 using the NASA-MSL and NASA-SMAP datasets because it can deal will long term dependencies but it is very resource-intensive. Isolation Forest had good performance on the KDD-TSAD data set (AUC-ROC = 0.93) while on high-dimensional data set it had relatively poor performance. ARIMA was also good with Linear trend on NAB Dataset having a highest AUC ROC of 0.85, but poor with non-Linear pattern. These findings reveal that particular algorithms are superior in definite situations but still, there is no algorithm that triumphs in handling all temporal patterns and varieties of anomalies.

The paper [3] presents various approaches of anomaly detection and the case study of Isolation Forest and Exponential Smoothing. These methods are quantitatively assessed by the authors in terms of factors like accuracy and the actual ranges from 85 to 99.12 percent. The study addresses important

obstacles such as parameter sensitivity and data distribution to stress the importance of such techniques, and the scarcity of resources needed for enhancing the detection's effectiveness in real-life settings.

The study [4] shows the The MSCRED model achieved an 85% F1-score on the Powerplant Dataset, demonstrating solid performance. However, the challenges remain significant. These include capturing temporal dependencies, encoding inter-correlations between different time series, ensuring robustness to noise, and providing varied anomaly scores based on the severity of incidents. Addressing these issues is crucial for improving the reliability and accuracy of anomaly detection in complex systems like powerplants.

B. Deep learning

The study [5] shows the CNN model applied to the OASIS dataset achieved an impressive precision of 94%. However, the evaluation lacked investigation into interpretability, causality, or uncertainty. These aspects are critical requirements in anomaly detection systems, as understanding the reasoning behind detections and considering causal relationships and uncertainties can enhance the reliability and usefulness of the model's output.The following study [6] shows OmniAnomaly, a solution tailored for enhancing the security of water treatment systems, demonstrated exceptional performance with an impressive F1-score of 99%. This achievement underscores its efficacy in detecting anomalies within critical infrastructure. Noteworthy is its emphasis on incorporating considerations of both order and causality in anomaly detection processes. By integrating these factors, OmniAnomaly significantly enhances its capacity to accurately identify deviations from anticipated behavior in water treatment systems, thus reinforcing their integrity and reliability.

The study [7] on anomaly detection highlights the importance of detecting unusual patterns in various fields such as cybersecurity, healthcare, industrial monitoring, and financial services. Swift and precise identification of anomalies allows organizations to proactively address potential catastrophic events, maintaining system integrity and operational continuity. Time series data, comprising sequences of values recorded over consecutive periods, requires analysis to recognize underlying patterns, trends, and seasonal variations. Fields like economics, meteorology, and engineering rely on time series analysis to predict future occurrences based on historical data. However, the complexity of time series data, which may exhibit non-stationarity, high frequency, and noise, presents challenges for analysis and forecasting. Combining anomaly detection with time series analysis enhances the ability to detect irregularities in time-dependent data. This combination is particularly critical in dynamic environments where data continuously evolves, and errors or detection delays can be costly. By employing advanced algorithms such as neural networks or machine learning models on time series data, analysts can pinpoint unusual patterns, predict potential disruptions, and take timely corrective actions. This methodology not only improves the precision of anomaly detection in complex datasets but also broadens its application across various sectors where predictive maintenance and real-time anomaly detection are essential.

In the following study [8], the DeepAnT model demonstrated strong performance on the Road Traffic Dataset, achieving an F1-score of 87%. Nonetheless, several challenges persist. Poor data quality and high contamination levels can compromise the modeling process by normalizing anomalies. Furthermore, selecting the appropriate network architecture and hyperparameters adds a significant layer of complexity to optimizing the model's performance.

The study [9] introduced us to VELC Model, a novel Variational Autoencoder model enhanced with LSTM for anomaly detection in time series data. On the KDD99 and Arrhythmia datasets for instance, VELC collects AUC scores of between 0. 722 to 0. 988, which is, really good compared to existing methods. The effectiveness of the model is also demonstrated in the study; however, the study also points out that the model is highly dependent on the chosen parameters and that the model is rather complex as the major weakness.

III. Methodology

This section delves into the proposed methodology highlighting the models' architectures utilized.

A. Preprocessing

The dataset is first checked for any missing values to ensure completeness. It is then sorted chronologically by the timestamp to maintain the correct temporal order. The timestamp is set as the index to facilitate time-based operations. The intervals between consecutive timestamps are calculated to ensure they are equidistant, identifying any irregularities. These steps clean and organize the data, preparing it for accurate time series analysis and subsequent modeling tasks.

B. Machine Learning Models

To detect anomalies in the power factor dataset, we employed three machine learning models: Isolation Forest, OneClass SVM, and Local Outlier Factor.Isolation Forest was used with a contamination rate of 0.02 and a random state of 42 to ensure reproducibility. This model succeeds in the isolation of observation by feature and random splitting values, which makes the type of model ideal in detecting anomalies in high dimensional spaces.OneClass SVM was configured with a (nu) parameter set to 0.02 and the gamma parameter set to 'auto'. The OneClass SVM finds a hyperplane that best separates the normal data from the origin in a transformed feature space, which helps in distinguishing outliers from regular data points.Local Outlier Factor (LOF) was implemented with 20 neighbors and a contamination rate of 0.02. The LOF model is particularly beneficial to anomalies since it looks at the dissimilarity of local density of a given point to their neighbors. It should however be noted that to be able to work with new observations, the novelty parameter is set to true.

We calculated the anomaly scores for each model, and the predictions were made to identify whether each data point

TABLE I
SUMMARY OF RELATED WORK ON AIR QUALITY PREDICTION

Ref	Dataset	Prediction Techniques	Evaluation Metrics	Limitations
[7]	Beijing PM2.5 Data	LSTM	RMSE, MAE	Requires large amounts of data, computationally intensive
[10]	Air Quality Monitoring Data	Deep Belief Networks (DBN)	Accuracy, Precision, Recall	Complex model, difficult to train
[5]	OASIS Dataset	CNN	94% Precision	Lack of Interpretability, Causality or Uncertainty were not investigated. requirements
[6]	Secure water treatment	OmniAnomaly	99% F1-score	Order and causality considerations.
[1]	Microsoft Dataset	SR-CNN	53.7% F1-score	Manual labeling and changing data make supervised anomaly detection insufficient.
[8]	Road Traffic Dataset	DeepAnT	87% F1-score	Poor data and high contamination compromise modeling, and network selection
[4]	Powerplant Dataset	MSCRED	85% F1-score	capturing temporal dependencies, encoding inter-correlations, ensuring noise robustness, and providing varied anomaly scores.
[2]	NASA-MSL, NASA-SMAP	LSTM-AD	AUC-ROC: 0.99%	High computational cost, memory-intensive
[3]	Various domains	Isolation Forest, Exponential Smoothing	Accuracy: 85%-99.12%	Sensitivity to parameters and data distribution
[9]	KDD99, Arrhythmia, ItalyPowerDemand, etc.	VELC (Variational AutoEncoder with LSTM)	AUC: 0.722-0.988	Sensitivity to parameter settings and model complexity

was an anomaly.These models were trained and tested on the 'INUT' feature of the dataset so that any peculiarities of the neutral line current was detected.

C. Deep Learning Models

The autoencoder uses Keras to define an autoencoder model that comprises of an encoder and a decoder. Starting with three Conv1D layers with progressively larger filter sizes (64, 32, and 16), the encoder section applies a relu activation function to each layer and uses'same' padding to preserve input form. The third convolutional layer's output is flattened, fed into a dense layer of 10 units, and then reconfigured to match the input shape of the decoder, which is (10, 1). The design of the encoder is reversed by the decoder, which begins with reshaping and goes on to include four Conv1D layers with progressively smaller filter sizes (16, 32, 64, and 128). Relu activation and "same" padding are used in each layer, and the final layer reconstructs the original input data using a single filter and a linear activation function. Often employed for applications like dimensionality reduction and anomaly detection, the goal of this architecture is to learn a compressed representation of the input data and then reconstruct the original input from this representation.

IV. EXPERIMENTAL RESULTS

The following section delves into the findings of two distinct types of experiments to provide a thorough evaluation of the proposed methodology. This analysis aims to offer a comprehensive understanding of the effectiveness of different models for anomaly detection.

A. Dataset Description

The dataset we used is called CurrentVoltage.csv. It contains power factor data from an AC power system. It consists of 19,352 rows and 10 features. The dataset has the following variables: VL1, VL2, VL3 (voltage measurements for three lines), IL1, IL2, IL3 (current measurements for three lines), VL12, VL23, VL31 (voltage measurements between lines), and INUT (neutral line current). These features collectively provide a comprehensive view of the power factor data, facilitating detailed analysis for anomaly detection.

B. Results

1) Experiment (1): Univariate Timeseries: The results indicate that the Isolation Forest model identified 375 instances of univariate anomalies, demonstrating a moderate level of sensitivity. In contrast, the One-class SVM detected a lower number of anomalies, with a total of 177 instances identified. The Local Outlier Factor model falls between the Isolation Forest and One-class SVM, detecting 265 univariate anomalies. Remarkably, the Autoencoder model detected only 1 univariate anomaly, suggesting either exceptional precision or limited sensitivity in identifying anomalies within this dataset.

Model	Univariate Anomalies
Isolation Forest	375
One-class SVM	177
Local Outlier Factor	265
Autoencoder	1

2) Experiment (2):Multivariate Timeseries: Across the models, the Isolation Forest, One-Class SVM, and Local

Outlier Factor models each identified a comparable number of multivariate anomalies, with counts of 388, 386, and 385 respectively. In stark contrast, the Autoencoder model exhibited a significantly higher detection rate, flagging a total of 1980 multivariate anomalies. This substantial disparity in detection rates between the Autoencoder and the other models suggests either a heightened sensitivity to anomalies or potentially a different approach to anomaly detection within multivariate timeseries data.

Model	Multivariate Anomalies
Isolation Forest	388
One-Class SVM	386
Local Outlier Factor	385
Autoencoder	1980

V. Conclusion

In conclusion, this research provides a thorough investigation into anomaly detection techniques within the realm of time series data analysis, particularly relevant to cybersecurity, health monitoring, and industry management domains. Through the meticulous exploration of machine learning and deep learning models, such as Isolation Forest, One-Class SVM, Local Outlier Factor, and autoencoder with CNN layers, this study sheds light on their efficacy in identifying anomalies across both single and multivariable datasets.

The significance of this inquiry lies in its contribution to advancing anomaly detection methodologies, offering researchers and practitioners a nuanced understanding of model performance in dynamic data environments. By conducting a comparative analysis using the Current Voltage dataset, this paper equips stakeholders with valuable insights to inform the selection of appropriate techniques tailored to specific dataset characteristics.

This work paves the way for further refinement and application of anomaly detection approaches, particularly in scenarios where timely detection is paramount for maintaining system integrity and operational continuity. Furthermore, the methodology outlined herein holds promise for adaptation and deployment across a spectrum of disciplines requiring predictive maintenance and real-time anomaly detection capabilities.

References

[1] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, "Time-series anomaly detection service at microsoft," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3009–3017, 2019.

[2] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proceedings of the VLDB Endowment*, vol. 15, pp. 1779–1797, 05 2022.

[3] D. Srivastava and L. Bhambhu, "Anomaly detection and time series analysis," 06 2023.

[4] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and 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, pp. 1409–1416, 2019.

[5] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection–a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–37, 2021.

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

[7] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proc. VLDB Endow.*, vol. 15, p. 1779–1797, may 2022.

[8] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "Deepant: A deep learning approach for unsupervised anomaly detection in time series," *Ieee Access*, vol. 7, pp. 1991–2005, 2018.

[9] Z. Chunkai and Y. Chen, "Time series anomaly detection with variational autoencoders," 07 2019.

[10] O. I. Provotar, Y. M. Linder, and M. M. Veres, "Unsupervised anomaly detection in time series using lstm-based autoencoders," in *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pp. 513–517, IEEE, 2019.