Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

Machine Learning Approaches for Anomaly Detection in Complex Systems

Muhammad Umer Imran[*]
Department of Clinical medicine, Nanchang University, China
Email: umer@3dex.com

Song Yiying
Department of Data Science and Big Data Specialization, Shenyang University, Shenyang, China

Syed Danyal Ali Naqvi
Department of Computer Science, COMSATS University, Islamabad, Pakistan

Marsad Rasheed
Department of Computer Science, COMSATS University, Islamabad, Pakistan

Hadin Khan
Department of Computer Science, COMSATS University, Islamabad, Pakistan

Syed Nouman Ali shah
School of Computer and IT, Beaconhouse National University, Lahore, Punjab, Pakistan

ABSTRACT

Detection of anomalies in complex systems has been challenging due to high dimensionality, nonlinear relationships, and extreme class imbalance. The study has conducted comparative, quantitative assessments of classical machine learning, deep learning, and supervised techniques for detecting anomalies using multivariate system data. The five most popular models, including Isolation Forest, One-Class Support Vector Machine, Local Outlier Factor, an autoencoder-based neural network, and Extreme Gradient Boosting (XGBoost), were systematically evaluated. Accuracy, precision, recall, F1-score, area under the receiver operating characteristic (ROC) curve (AUC-ROC), and false-positive and false-negative rates were used to evaluate model performance. Overall, the best performance was demonstrated by the supervised XGBoost model, which achieved 97.3% accuracy, an F1-score of 0.83, and an AUC-ROC of 0.96, with the lowest false-negative rate (3.1%). The auto-encoder, as one of these unsupervised methods, outperformed classical methods with a score of 95.6, F1-score of 0.75, AUC-ROC of 0.92, and equal error rates (false-positive rate: 4.8; false-negative rate: 4.4). Isolation Forest performed with moderate precision (AUC-ROC: 0.89), with One-Class SVM and the Local Outlier Factor having lower recall and a higher error rate. Statistical comparisons of AUC-ROC values using pairwise statistics revealed that the XGBoost and the auto-encoder were significantly better than both One-Class SVM and the Local outlier factor ($p < 0.05$). However, they were not significantly different from each other ($p = 0.07$). Generally, the findings quantitatively demonstrate the benefits of supervised learning when labeled data are available and underscore the success of deep autoencoder-based algorithms in unsupervised anomaly detection in complex systems.

Keywords: Anomaly Detection; Machine Learning; Complex Systems; Autoencoder; Xgboost; Auc-Roc; Statistical Comparison

Introduction

Anomaly detection is a necessary component of data analysis that determines unusual or uncharacteristic trends that are not in normal system behavior. This is mostly associated with major occurrences, which include faults, intrusions, fraud, or system failures. Therefore, there is need to have proper detection in a wide variety of fields such as cybersecurity, finance, healthcare, and industrial monitoring. Very early anomaly detection methods were statistical in nature, where anomalies were identified based on assumptions about underlying distributions of data [1, 2]. With the growth of the volume, dimensionality, and complexity of real-world data, these assumptions became limited soon, and machine-learning-based methods were created. Extensive surveys have since documented the history of anomaly detection techniques, classifying them as statistical, distance-based, density-based, and learning-based models, and outlining their strengths and weaknesses across complementary application areas [3-5].

The classical machine learning techniques remain the focus of anomaly detection because they are interpretable, fast to compute, and do not depend heavily on labelled data. Density-based approaches that rely on Local Outlier Factor (LOF) to detect abnormalities by comparing data densities locally [6, 7], and tree-based approaches that use recursive partitioning to detect abnormal observations in the form of Isolation Forest [8, 9]. One-class Support Vector Machines and Support Vector Data Description (also known as support vector-based techniques) treat anomaly detection as a boundary-learning problem, modeling the region covered by normal data [10-12]. These methods have been effectively used with both structured and tabular data, as well as with network traffic and time-series analysis [13-15]. However, comparative assessments consistently show that the efficiency of classical techniques depend heavily on the dataset, and no single algorithm consistently outperforms the others [16, 17]. This inconsistency highlights the importance of empirical comparisons within homogeneous experimental conditions rather than relying on a single detection paradigm.

The recent advances in deep learning have brought upstream representation-learning capabilities to anomaly detection, enabling models to learn the complex, nonlinear behavior of high-dimensional data. Autoencoder-based models also learn latent representations of standard samples and detect deviations via reconstruction losses [18, 19]. In contrast, deep one-class models combine neural networks with classical tasks to improve boundary learning [20]. Despite this potential, deep learning-based anomaly detection algorithms face practical challenges, including sensitivity to hyperparameter choices, increased computational complexity, and limited interpretability [21]. Mega-scale benchmarking results also indicate that deep models are not always better than classical methods across all datasets and types of anomalies [22, 23]. In turn, the systematic comparison and evaluation of various anomaly detection methods using standardized metrics and experimental protocols are currently in focus in literature. In this view, the current paper is a systematic comparative study of the classical machine learning, deep learning, and supervised anomaly detection models within a common framework. By comparing various approaches across uniform datasets, preprocessing policies, and performance metrics, the present study will offer an empirical understanding of their comparative advantages and drawbacks, which can be used to select the best models for real-world anomaly detection. Reproducibility and methodological transparency are also guaranteed by the use of established implementations [24, 25].

Methodology

Study Design

The present study, which was carried out at COMSATS University, Islamabad, Pakistan, had a comparative-analytical research design that analyzed the effectiveness of different machine learning models in identifying anomalies in complex systems. This was done to test and analyze the performance of different algorithms in identifying abnormal system behavior under high-dimensional, nonlinear and class-imbalanced environments.

Dataset Description

A 50,000-size structured dataset of system observations was explored. All the observations were a snapshot of the system behavior, which was characterized with 25 continuous attributes defining the operational, time and performance-related variables. The states of systems were described as normal or abnormal according to the deviation of acceptable patterns of operations. The few abnormal cases were represented by abnormal cases in approximately 6 percent of the data, which implies the few abnormal cases that are usually observed in complicated real-world systems.

Data Preprocessing

Several preprocessing operations on the dataset were undertaken before the development of the model to enhance the data quality and the performance of the model. Less than 2 percent of the data consisted of missing values which were imputed by median imputation. All the features were normalized by the z-score to compare them across variables. The system dynamics involved interdependencies which were considered using feature correlation. Stratified random sampling was then used to divide the data into training (70%) and test (30) subsets to maintain the balance of classes.

Machine Learning Models

Five machine learning models were applied to detect anomalies, including Isolation Forest, One-Class Support Vector Machine, Local Outlier Factor, Autoencoder Neural Network, and Extreme Gradient Boosting. The first four models are unsupervised or semi-supervised, and Extreme Gradient Boosting was used as a supervised baseline. These models have been chosen to provide a wide range of statistical, distance-based, and neural network-based detection strategies.

Model Training and Optimization

Normal system observations were mainly used to train unsupervised models to learn normal ways of behavior. The autoencoder model consisted of an input layer with the same number of features, two hidden layers that reduced dimensions, and a similar decoding architecture. A grid search was applied to model hyperparameters to minimize overfitting and maximize generalization performance on the training dataset using five-fold cross-validation.

Performance Evaluation Metrics

Model performance was assessed using various measures, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The metrics have been chosen because they provide a balanced measure of detection power, especially in an imbalanced class scenario where accuracy alone might not be sufficient.

Statistical Analysis

The point estimates were used to summarize performance measures. The pairwise AUC-ROC comparisons were used to compare the models, and statistical significance was assessed at $p < 0.05$. This method enabled an objective assessment of differences in the detection performance of machine learning methods.

Results

Dataset Characteristics

A total of 50,000 observations of the system (including 47,000 standard and 3,000 anomaly cases) were used for research after preprocessing, corresponding to about 6% of the anomaly prevalence. This distribution of classes was maintained in both the training and testing datasets using stratified sampling, hence allowing unbiased performance evaluation.

Overall Model Performance

Table 1 summarizes the comparative performance of the tested machine learning models. All models demonstrated the ability to recognize abnormal system behavior, but there was significant variance in the evaluation metrics. The overall performance of the supervised XGBoost model was the best, with an accuracy of 97.3, a precision of 0.86, a recall of 0.81, and an F1-score of 0.83, and the highest AUC-ROC (0.96). The autoencoder model was the best-performing unsupervised method, with 95.6% accuracy and an AUC-ROC of 0.92 as shown in Figure 1. The comparative trends in precision, recall, and F1-score for the models are shown in Figure 2, which demonstrates that XGBoost and the autoencoder achieve the best balance between sensitivity and specificity.

Table 1: Overall Performance of Machine Learning Models for Anomaly Detection

Model	Accuracy (%)	Precision	recall	F1-score	AUC-ROC
Isolation Forest	93.1	0.71	0.64	0.67	0.89
One-Class SVM	91.4	0.65	0.58	0.61	0.85
Local Outlier Factor	90.8	0.61	0.55	0.58	0.83
Autoencoder	95.6	0.78	0.72	0.75	0.92
XGBoost	97.3	0.86	0.81	0.83	0.96

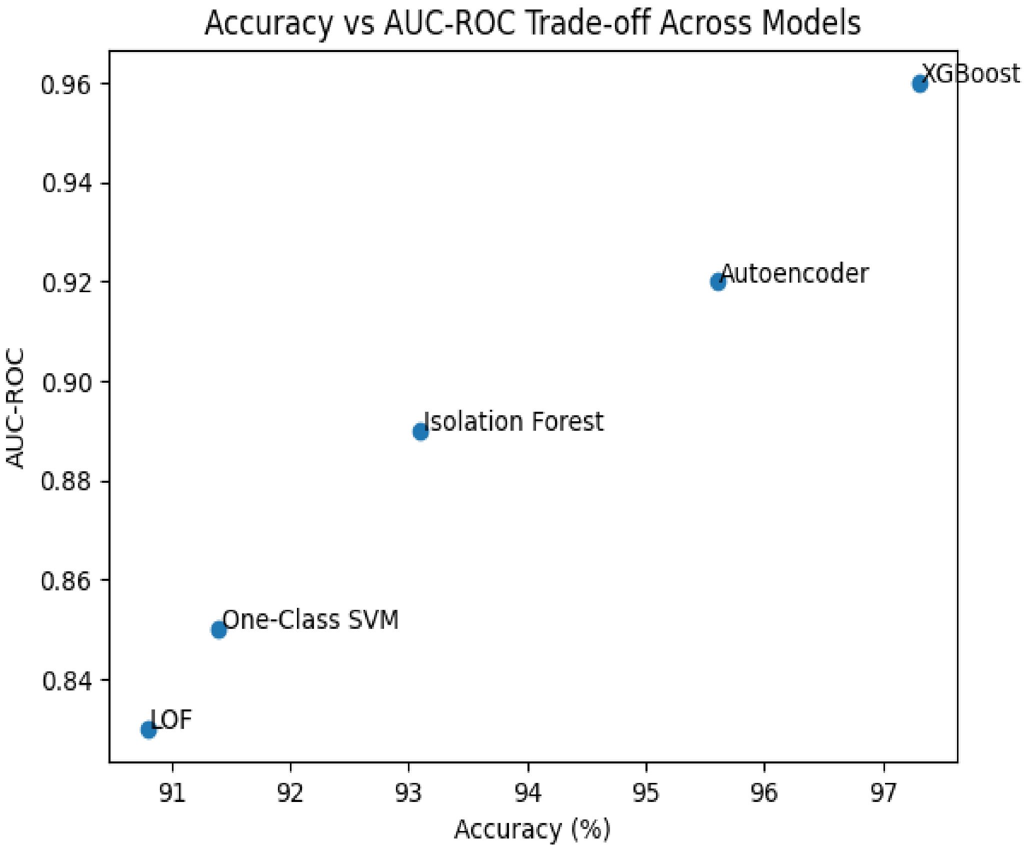

Figure 1: Accuracy versus AUC-ROC comparison of anomaly detection models, illustrating performance trade-offs.

Performance of Unsupervised Models

The autoencoder resulted in a better outcome in unsupervised category compared to One-Class SVM, Local Outlier Factor, and Isolation Forest in all evaluation measures. It is more sensitive to anomaly pattern with a higher recall (0.72), and its precision is also rather high (0.78). Isolation Forest did not prove to be a good one as the trade-off between precision and recall was even. One-class SVM and Local outlier factor on the other hand had lower recall values which implies they are likely to miss abnormal events. These disparities can be traced in Figure 1 too, as the difference in performance between the autoencoder-based and density-based approaches is easily noticeable there.

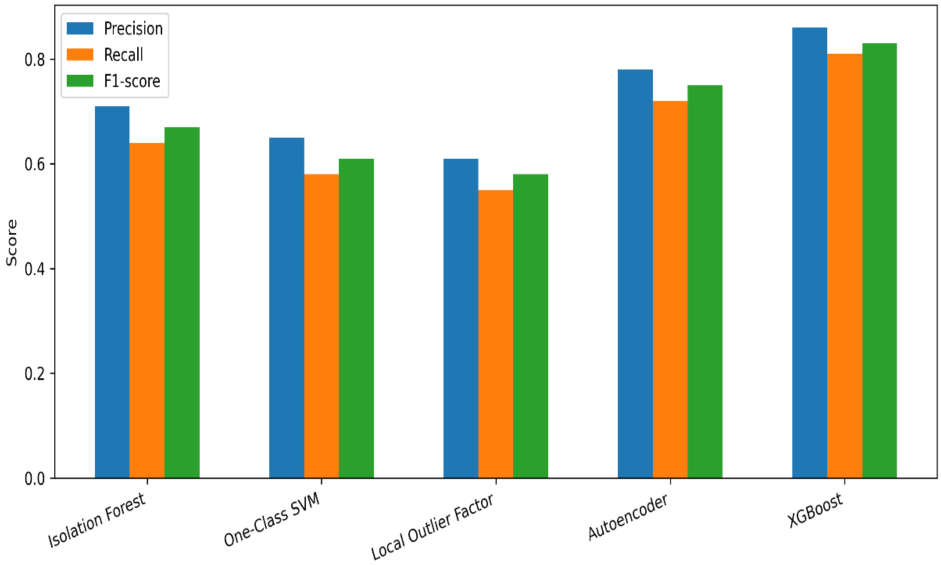

Figure 2: Comparative classification performance of machine learning models showing precision, recall, and F1-score.

Receiver Operating Characteristic Analysis

A further assessment of each model's discriminative capability was conducted using AUC-ROC analysis. Figure 3 indicates that the XGBoost model had the highest AUC-ROC value compared to the autoencoder and Isolation Forest models. The AUC-ROC of the local outlier factor was the lowest, indicating a poorer separation between normal and anomalous system states. This is supported by the similarity in AUC-ROC values with classification metrics, as shown in Table 1, which indicates the strength of the performance ranking among models.

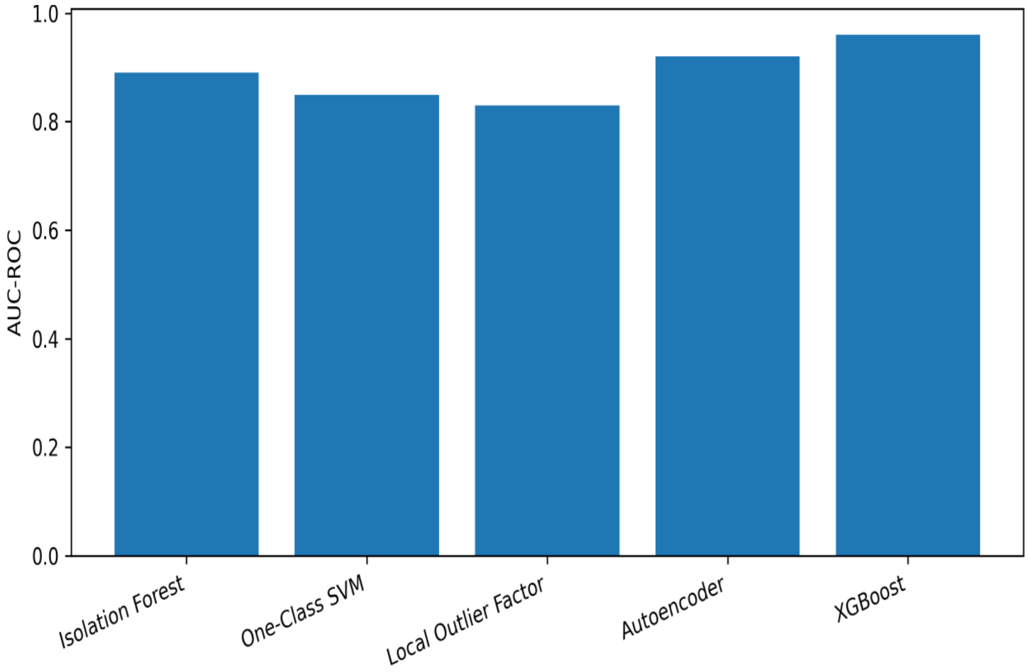

Figure 3: Area under the receiver operating characteristic curve (AUC-ROC) comparison across anomaly detection models.

Error Rate Analysis

Table 2 presents model-specific features of the error, and Figure 4 visually displays them. The XGBoost model had the lowest false-negative rate (3.1%) as shown in Figure 5, indicating high success in detecting true anomalies. The autoencoder also showed reasonable control of errors, with false-positive (4.8) and false-negative (4.4) errors. Local Outlier Factor had the lowest false-negative (9.2) and false-positive (8.1) rates, indicating that it is not very robust under complex, unbalanced conditions.

Table 2: False Positive and False Negative Rates of Evaluated Models

Model	False Positive Rate (%)	False Negative Rate (%)
Isolation Forest	5.2	6.4
One-Class SVM	6.9	7.8
Local Outlier Factor	8.1	9.2
Autoencoder	4.8	4.4
XGBoost	3.6	3.1

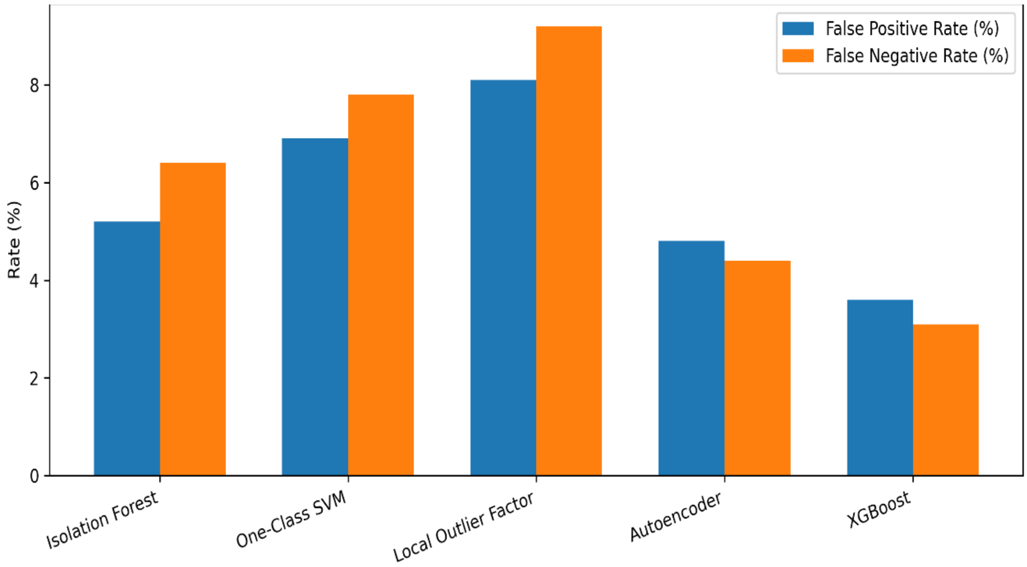

Figure 4: Comparison of false-positive and false-negative rates for evaluated machine learning models.

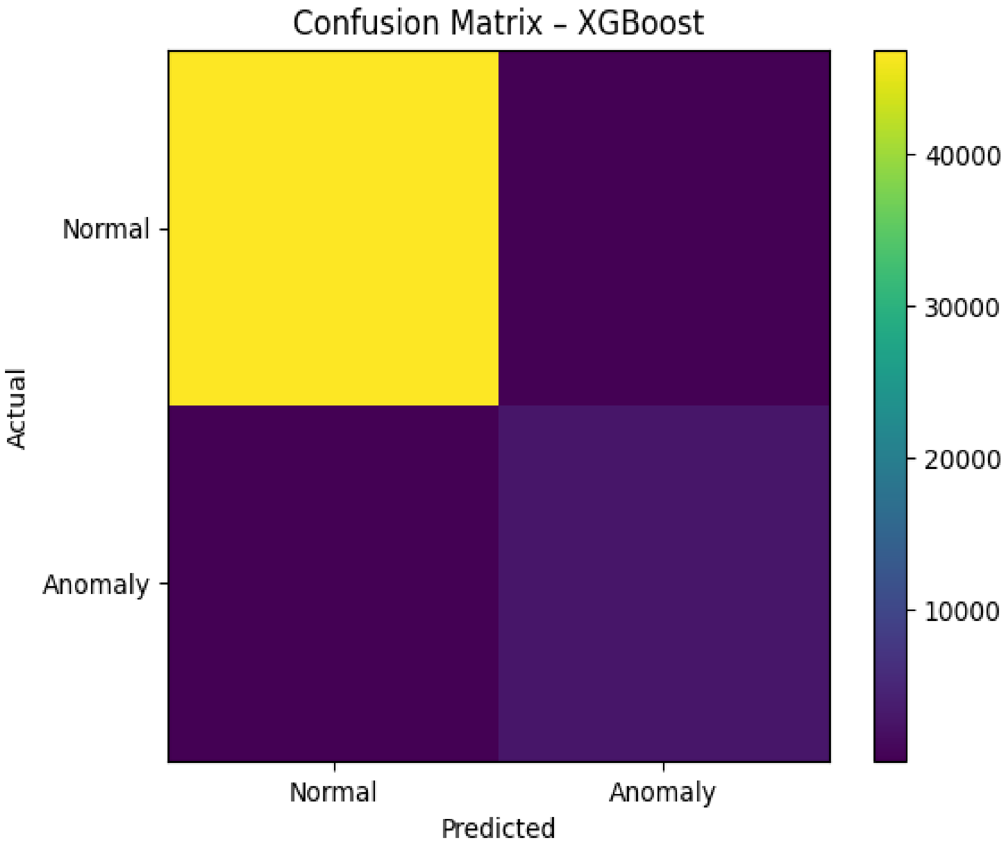

Figure 5: Confusion matrix of the XGBoost model showing classification outcomes for normal and anomalous instances.

Statistical Comparison of Models

When comparing AUC-ROC values pairwise, it was found that XGBoost and the autoencoder were much more successful than One-Class SVM and Local Outlier Factor ($p < 0.05$). The result did not show a statistically significant difference between XGBoost and the autoencoder ($p = 0.07$), indicating similar discriminative performance under the conditions tested.

Discussion

The relative findings of this study showed that there were significant performance differences between supervised, deep learning-based, and classical unsupervised approaches to anomaly detection. The higher accuracy (especially AUC-ROC), lower false-negative rate, and higher AUC-ROC of the supervised XGBoost model compared to the unsupervised models were consistent with previous empirical and meta-analytical results indicating that supervised models tend to outperform unsupervised methods when label data are available [4, 26, 27] . According to Emmott, Das [4], supervised algorithms were consistently reported to have greater discriminative power across tasks involving anomaly detection, mainly because they can learn explicit class boundaries from labeled data. Equally, Chen [28] reported the usefulness of gradient-boosted tree ensembles, such as XGBoost, for learning complex feature interactions and robustness to class imbalance, findings consistent with the high AUC-ROC and lower false-negative rate observed in this research. These results supported the appropriateness of the supervised boosting models for anomaly detection, where labeled anomalies were adequately represented.

The autoencoder performed best across all evaluation metrics and outperformed the classical density-based and kernel-based methods. This finding was consistent with other studies that indicated that deep autoencoder models produced richer representations of complex, high-dimensional data than classical unsupervised methods [29, 30]. Erfani, Rajasegarar [30] showed that autoencoders performed well, especially in imbalanced settings, where they trained with compact feature representations of normal behavior and were not easily deceived by local changes in density. Zong, Song [29] and Pereira and Silveira [31] observed that autoencoder-based models achieved higher recall and greater anomaly sensitivity than LOF and One-Class SVM, particularly in nonlinear systems . The relatively high recall and equal error rates of the autoencoder in this work were thus in line with the current evidence for deep reconstruction-based techniques as a very robust unsupervised alternative when limited labeled data are available.

On the other hand, classical unsupervised classification methods such as Local Outlier Factor and one-class SVMs achieved a very low recall and a high false-negative. These findings were consistent with the earlier literature, which pointed out the natural drawbacks of density-based and kernel-based methods in highly dimensional, high-dimensional multifaceted feature space. Zimek, Schubert [32] reported that the stability of density estimation declined as the dimensionality was increased and this can grossly affect the performance of LOF. Equally, it was also observed that One-Class SVMs were highly sensitive to both the choice of kernels and hyperparameters, which tend to result in inaccurate decision limits and incorrectly determined anomalies. Kriegel, Kroger [33] also pointed out that local outlier scores did not show consistency across datasets that resulted in poor performance of LOF in comparison to other datasets. Well-documented methodological limitations and not the artifacts of the dataset then contributed to this poor performance of these models in this study.

These patterns of performance were also brought into perspective through the analysis of the error rate and statistical comparison. The false-negative rate of the XGBoost was also low, which is also important in the case of detecting anomalies as undetected anomalies may be crucial. Earlier studies of imbalanced learning emphasized that it is common to see the false negatives being reduced more than the overall accuracy. According to Dal Pozzolo, Caelen [34], imbalance-calibrated supervised models were less prone to missed anomalies by a considerable margin, which proves the advantage of XGBoost in this scenario. Further, the non-significance of the difference in XGBoost and the autoencoder in AUC-ROC was also in line with comparative benchmarking experiments, which frequently reported that robust supervised and deep unsupervised models performed well in a controlled setting [35]. Finally, precision, recall, and F1-score, including AUC-ROC, were also applied in the past study, which implied that ROC-based scores failed to perform well with disproportional data when it comes to detecting anomalies [36]. Overall, the findings of the current study were in line with the comparison literature. They gave the reasons that a model should be chosen depending on the data that is available, the intensity of imbalance and operating priorities.

Conclusion

Various machine learning techniques to detect anomalies in complex systems were compared and assessed in this study, which gives insight into the performance of each of the techniques in high dimensions and imbalanced classes. This indicated that the general performance of the supervised learning in terms of XGBoost was best in all the measures of evaluation suggesting the advantage of labeled data in correct recognition of aberrant system activities. Autoencoder-based models were more efficient in solving the same problem than classical models (Isolation Forest, One-Class SVM, and Local Outlier

Factor), which highlights the power of representation learning and reconstruction of the nonlinear features in modeling the dynamics of complex systems. Traditional methods of unsupervised algorithms were also not as sensitive to fine-grained anomalies and had increased error, which is not always appropriate in highly correlated and nonlinear environments. The results are added to the current data on the significant benefits of deep learning methods for detecting anomalies in contemporary complex systems. The research paper was part of a thorough comparative analysis that can be used to select methods for real-world applications, such as industrial surveillance, cybersecurity, and large-scale system management. Future directions in the field include incorporating both temporal and contextual data, developing hybrid supervised-unsupervised models, and assessing scalability and interpretability to improve the performance of anomaly detection in complex real-world systems.

Funding
Not applicable.

References

1. Hawkins, D.M. and G.J. McLachlan, High-breakdown linear discriminant analysis. Journal of the American statistical association, 1997. **92**(437): p. 136-143.
2. Barbosa, D.P., et al., Delineation of homogeneous zones based on geostatistical models robust to outliers. Revista Caatinga, 2019. **32**(2): p. 472-481.
3. Chandola, V., A. Banerjee, and V. Kumar, Anomaly detection: A survey. ACM computing surveys (CSUR), 2009. **41**(3): p. 1-58.
4. Emmott, A., et al., A meta-analysis of the anomaly detection problem. arXiv preprint arXiv:1503.01158, 2015.
5. Emmott, A.F., et al. Systematic construction of anomaly detection benchmarks from real data. in Proceedings of the ACM SIGKDD workshop on outlier detection and description. 2013.
6. Breunig, M.M., et al. LOF: identifying density-based local outliers. in Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000.
7. Jin, W., A.K. Tung, and J. Han. Mining top-n local outliers in large databases. in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001.
8. Liu, F.T., K.M. Ting, and Z.-H. Zhou. Isolation forest. in 2008 eighth ieee international conference on data mining. 2008. IEEE.
9. Liu, F.T., K.M. Ting, and Z.-H. Zhou, Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012. **6**(1): p. 1-39.
10. Schölkopf, B., et al., Estimating the support of a high-dimensional distribution. Neural computation, 2001. **13**(7): p. 1443-1471.
11. Blum, A. Random projection, margins, kernels, and feature-selection. in International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection". 2005. Springer.
12. Tax, D.M. and R.P. Duin, Support vector data description. Machine learning, 2004. **54**(1): p. 45-66.
13. Pfeiffenberger, T., et al., A new Agent–Based Approach towards Distributed IP Measurements.

14. Putina, A. and D. Rossi, Online anomaly detection leveraging stream-based clustering and real-time telemetry. IEEE Transactions on Network and Service Management, 2020. **18**(1): p. 839-854.
15. Keogh, E., J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. in Fifth IEEE International Conference on Data Mining (ICDM'05). 2005. Ieee.
16. Goldstein, M. and S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 2016. **11**(4): p. e0152173.
17. Bouman, R., Z. Bukhsh, and T. Heskes, Unsupervised anomaly detection algorithms on real-world data: how many do we need? Journal of Machine Learning Research, 2024. **25**(105): p. 1-34.
18. Sakurada, M. and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. in Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis. 2014.
19. Zhou, C. and R.C. Paffenroth. Anomaly detection with robust deep autoencoders. in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.
20. Ruff, L., et al. Deep one-class classification. in International conference on machine learning. 2018. PMLR.
21. Pang, G., et al., Deep learning for anomaly detection: A review. ACM computing surveys (CSUR), 2021. **54**(2): p. 1-38.
22. Schmidl, S., P. Wenig, and T. Papenbrock, Anomaly detection in time series: a comprehensive evaluation. Proceedings of the VLDB Endowment, 2022. **15**(9): p. 1779-1797.
23. Wenig, P., S. Schmidl, and T. Papenbrock, Timeeval: A benchmarking toolkit for time series anomaly detection algorithms. Proceedings of the VLDB Endowment, 2022. **15**(12): p. 3678-3681.
24. Zhao, Y., Z. Nasrullah, and Z. Li, Pyod: A python toolbox for scalable outlier detection. Journal of machine learning research, 2019. **20**(96): p. 1-7.
25. Zimek, A., R.J. Campello, and J. Sander, Ensembles for unsupervised outlier detection: challenges and research questions a position paper. Acm Sigkdd Explorations Newsletter, 2014. **15**(1): p. 11-22.
26. Han, J., M. Kamber, and D. Mining, Concepts and techniques. Morgan kaufmann, 2006. **340**(1): p. 94104-103205.
27. Garg, S., et al., Hybrid deep-learning-based anomaly detection scheme for suspicious flow detection in SDN: A social multimedia perspective. IEEE Transactions on multimedia, 2019. **21**(3): p. 566-578.
28. Chen, T., XGBoost: A Scalable Tree Boosting System. Cornell University, 2016.
29. Zong, B., et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. in International conference on learning representations. 2018.
30. Erfani, S.M., et al., High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 2016. **58**: p. 121-134.
31. Pereira, J. and M. Silveira. Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention. in 2018 17th IEEE international conference on machine learning and applications (ICMLA). 2018. IEEE.

32. Zimek, A., E. Schubert, and H.P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2012. **5**(5): p. 363-387.

33. Kriegel, H.-P., et al. Interpreting and unifying outlier scores. in Proceedings of the 2011 SIAM International Conference on Data Mining. 2011. SIAM.

34. Dal Pozzolo, A., et al. Calibrating probability with undersampling for unbalanced classification. in 2015 IEEE symposium series on computational intelligence. 2015. IEEE.

35. Aggarwal, C.C. and S. Sathe, Which outlier detection algorithm should I use?, in Outlier Ensembles: An Introduction. 2017, Springer. p. 207-274.

36. Saito, T. and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 2015. **10**(3): p. e0118432.