International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 1993-2017
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

A Hybrid Framework for Robust Anomaly Detection: Integrating Unsupervised and Supervised Learning with Advanced Feature Engineering

Girish Reddy Ginni[1]*, Srinivasa L. Chakravarthy[2]

[1]Department of Computer Science Engineering, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh-530 045.
* **Corresponding Author Email:** girishloshankar@gmail.com - **ORCID:** 0009-0005-5242-8839

[2]Department of Computer Science and Engineering, College GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh-530 045.
Email: chakri.ls@gmail.com- **ORCID:** 0000-0001-9141-4863

Article Info:

DOI: 10.22399/ijcesen.1383
Received : 03 January 2025
Accepted : 16 March 2025

Keywords :

Hybrid Anomaly Detection,
Unsupervised Learning,
Supervised Learning,
Feature Engineering,
Outlier Detection.

Abstract:

Finding anomalous data is essential in various applications, from cyber security to healthcare to industrial monitoring. Traditional methods- unsupervised or supervised—are far from straightforward; unsupervised methods are notoriously plagued by high false favorable rates and unclear distinction boundaries, while supervised methods tend to rely on a great deal of labeled data, often in limited supply or highly imbalanced. Indeed, these problems call for a unified approach that takes advantage of the benefits of both paradigms for more robust anomaly detection. In this work, we develop a hybrid outlier detection framework combining several unsupervised anomaly scoring models (Isolation Forest, Local Outlier Factor, and One-Class SVM) and XGBoost and Logistic Regression as a supervised classifier. Instead, we combine the proposed algorithm with advanced feature engineering techniques (e.g., topological space optimization) to extract informative features for our data representation. Our empirical studies of diverse benchmark datasets (Arrhythmia, Cardio, Letter, Mammography, MNIST, Satellite, and Speech) indicate that the hybrid model consistently shows a significant improvement over any single method. Our framework consistently reduces false positives and false negatives and is more precise; recall, F1-score, and ROC-AUC are the highest scores for quantitative comparison. We demonstrate the usefulness of the proposed framework by enabling it to handle high-dimensional, imbalanced datasets while leading to meaningful detection results in real-world settings. Establishes a new state-of-the-art performance in anomaly detection while also supplying an approach that is scalable and versatile for complex data environments and forming a basis from which to build toward future integrated anomaly detection systems.

1. Introduction

Due to the importance of detecting rare events or outliers from high-dimensional data in maintaining system availability and decision-making, anomaly detection has become a key research topic in many application areas, including Cybersecurity, medical diagnostics, and remote sensing. Conventional approaches are primarily based on unsupervised or supervised methods, each with drawbacks. For example, unsupervised methodologies like Isolation Forest and Local Outlier Factor can suffer from high false positives [1,2]. At the same time, supervised methods like XGBoost and Logistic Regression rely on labeled data, which is often unavailable in the real world. Recently, this idea has attracted the attention of researchers in the community and has led to the design of hybrid methods that combine the advantages of both paradigms to obtain a more reliable detection performance [3,4].

To overcome these challenges, the present research proposes a new hybrid outlier detection algorithm that shows a novel integration of unsupervised anomaly scoring with supervised classification and effective feature engineering techniques. Thus, this research aims to propose an integrated framework that combines the attractive advantages of various models, thereby increasing detection accuracy and decreasing the number of false positive and false

negative results. This approach features a novel dynamic feature selection paradigm based on topological subspace optimization. This refined ensemble strategy combines anomaly scores of different unsupervised methods and statistical performance measures to validate robustness. The research contributions are extensive: the proposed model is assessed on a variety of benchmark datasets, which includes Arrhythmia, Cardio, Letter, Mammography, MNIST, Satellite, and Speech, achieving better performance than the state-of-the-art methods. Furthermore, the hybrid approach dramatically improves detection with extensive experiments and statistical analyses and helps gain valuable insights for practical implementations.

We summarize this paper as follows. In Section 2, we provide a comprehensive literature survey reviewing related sources and existing methods for anomaly detection methods, leading to the conclusion of remaining gaps in current functionality. In Section 3, we propose our methodology by explaining how we combine unsupervised and supervised models with feature engineering. Experimental results with performance on several datasets using precision, recall, F1-score, and ROC-AUC are in Section 4. Section 5 presents the implications of the findings and limitations of the present study, and Section 6 ends the paper and provides future research directions. Enabling this structured methodology, the Research tries to prove the gap between existing methods and to conclude a more Policy-level solution and Scalable Anomaly Detection Framework.

2. Related Work

This survey of contemporary methods for hybrid and deep learning enhances a literature review of more hybrid and deep learning-based anomaly detection methods, identifying gaps in the current addiction literature and positioning a need for developing our framework. Jeffrey et al. In [1], it introduces a hybrid anomaly detection model for Cyber-Physical Systems (CPS) based on a mixture of machine learning, threshold-based & signature-based techniques to increase the efficiency and accuracy of threat detection. Yaro et al. For outlier detection, they compared three hybrid scale estimators (weighted, maximum, and average) for the mZ-score method and found that the weighted hybrid approach is the most successful [2].HVK/HVA-HEM hybrid framework for river discharge prediction combines several models to get better results than current approaches and can potentially be improved even further by incorporating climatic factors. Hu et al. [4] ASOD technology improves online anomaly detection in stream data by employing adaptive algorithms and dynamic context management, surpassing previous approaches regarding accuracy and stability. Future research will concentrate on enhancing multi-variate time series identification and developing a more profound comprehension of abnormalities. Stehle et al. [5] provided scalable anomaly detection on HPC clusters; DeepHYDRA combines deep learning and DBSCAN. This allows for real-time detection with minimal resource consumption and addresses dimensionality concerns. Future goals will include further customization and optimization for different system settings.

Alsmadi et al. [6] The TCLD method solves outlier and topic count problems to improve topic modeling for short texts. It performed better in clustering, but complicated data requires tuning. Nssibi et al. [7], the iBABC-CGO method improves gene selection and increases accuracy and efficiency in high-dimensional datasets by combining algorithms for chaotic game optimization with artificial bee colonies. Future work aims to enhance initialization and apply it to more domains. Gouranga and Rajiv [8] suggested that the hybrid approach raises the AUC score of contextual outlier identification by 22–45% by fusing neural networks with conventional methods. Adaptive ensemble learning will be a part of future work. Xie and Huang [9] proposed a hybrid sampling technique that uses Mahalanobis distance SMOTE-ENN and Random Forest to identify credit card fraud. This approach enhances accuracy by 22–45%. Future studies will concentrate on neural network integration for broader applications. Jiang et al. [10] improved river pollution identification; the study creates a hybrid anomaly detection framework that combines SVDD and VMD-BPNN. It performs better than alternative approaches, and future research suggests using deep learning for more accurate forecasts.

Princz et al. [11] compared and contrasted several machine learning models for investigating anomaly detection in binary time series. Future studies will improve the size of the dataset, the thresholds, and the immediate execution. Alghushairy et al. [12] introduced an improved anomaly-based GNB and SVM-based network outlier detection system (NODS). Future research will examine deep learning approaches to enhance detection abilities. Ferreira et al. [13] evaluated supervised and unsupervised machine learning techniques for textile fault identification using autoencoders. Subsequent investigations will focus on improving unsupervised methods and evaluating substitute approaches. Montalvo et al. [14] offered a hybrid anomaly detection technique for Internet of Things

devices to decrease bandwidth usage and increase security. Additionally, it seeks to enhance and better analyze the data. Ali et al. [15] enhanced reservoir characterization and reduced expenses by reconstructing density logs from irregular healthy data through supervised and unsupervised machine learning techniques. Projects, including regional applications and seismic integration, are coming up soon.

Marques et al. [16] evaluated unsupervised outlier identification techniques against one-class classification to discover anomalies. Ensembles exhibit higher accuracy compared to both SVDD and GMM. We will continue to research combination tactics and variety in ensembles. Cui et al. [17] highlighted advances in unsupervised anomaly detection for Industry 4.0, stressing the benefits of deep learning over traditional methods and its data and efficiency constraints. Subsequent studies ought to enhance the generality of models and address dataset constraints. Islam et al. [18] proposed that the Credit Card Anomaly Detection (CCAD) model outperforms traditional methods in detecting anomalies in minority groups. Future work will explore deep learning methods using CCAD on several datasets. Samunnisa et al. [19] unique hybrid intrusion detection system effectively categorizes abnormalities associated with cloud computing. More research will enhance modeling methodologies to enhance detection accuracy and manage dynamic assault patterns. Kennedy et al. [20] enhanced this by adding an unsupervised fraud detection method that considers label accuracy and class imbalance, surpassing existing models. Additional studies will look at more datasets and baseline learners.

Dash et al. [21] suggested an IQR-based winsorizing procedure for detecting outliers, which was followed by a TLBO-based model categorization of the data. Further research aims to investigate big data applications and enhance TLBO attributes. Koko et al. [22] suggested a method for identifying outliers in clustering-based dynamic selection (CBDS) that performs more accurately and efficiently than LSCP. Future research will examine more advanced grouping techniques and enhanced fusion techniques. Haque et al. [23] examined the benefits and limitations of using machine learning algorithms for Wireless Sensor Network (WSN) anomaly detection. Future studies will concentrate on large-scale network algorithm optimization and real-world assessment. Lee et al. [24] introduced a hybrid deep learning model that offers benefits over earlier techniques for anomaly identification in smart factories. Managing noisy data and improving model interpretability should be the main areas of future study. Savic et al. [25] proposed HUNOD, a hybrid unsupervised method that enhances tax fraud detection accuracy and interpretability by fusing clustering with representation learning. Future studies should increase the accuracy of labels.

Velasquez et al. [26] formed a hybrid machine learning ensemble using LOF, OCSVM, and Autoencoder for Industry 4.0 real-time anomaly identification. Prospective investigations should focus on deep learning, retraining costs, and different fault classifications. Zheng et al. [27] proposed a hybrid method that blends deep neural networks with hyperspheres for high-dimensional anomaly detection. Future studies in similar settings will address the interference caused by unrelated features. Sakhnenko et al. [28] presented a hybrid classical-quantum autoencoder (HAE) for anomaly detection that combines quantum and classical models for improved performance. We will look at hybrid models and measuring techniques in further research. Karitonov et al. [29] evaluated eleven machine learning models, with KNN proving to be the most effective for manufacturing anomaly detection. Plans include collecting more data and analyzing previous logs. Fazlic et al. [30] introduced a hybrid anomaly detection method that integrates statistics, SOM, and LDA for real-time medical data. Stochastic Petri nets and genetic algorithms combined for optimization will be investigated further. Chander and Kumaravelan [31] examined the latest methods for detecting outliers in Wireless Sensor Networks (WSNs) while emphasizing the difficulties related to bandwidth, computing, and energy. Future research will focus on enhancing detection accuracy and addressing dataset variability. Zhou et al. [32] presented HAD-MDGAT, a hybrid anomaly detection model that combines MDA and GAT for enhanced multivariate time series analysis. The strategy improves accuracy over current approaches by addressing correlations in both space and time. Stability in GAN training and sliding window optimization will be investigated in further study. Thudumu et al. [33] looked at high-dimensional big data anomaly detection problems, emphasizing the shortcomings of existing methods and the need for new frameworks to improve performance and accuracy. These problems should be addressed in future research. Wang and Mao [34] proposed an ensemble-based approach to identify outliers in industrial systems without labeled data. One-class and multi-class classifiers are integrated to address real-world problems and the need for more research. Kurt et al. [35] proposed scalable and nonparametric methods for real-time anomaly detection in high-dimensional data. It addresses the shortcomings of traditional methods, including the

handling of non-stationary data. Song et al. [36] suggested that the three-phase HFS-C-P algorithm effectively handles high-dimensional issues by integrating several feature selection techniques. Reducing processing costs and improving clustering approaches should be the main goals of future development. Qaraad et al. [37] compared to current methods, the ENSVM model identifies appropriate gene subsets for cancer classification effectively. The interpretation of gene importance for cancer therapy will be the main focus of future research. Yuan et al. [38] presented an outlier identification technique (FIEOD) based on fuzzy information entropy (FIEOD), which outperforms classical methods on a variety of data types. Dynamic outlier identification will be investigated in future studies. Chen et al. [39] recommended that the LPPCA method improves high-dimensional anomaly detection by employing Locally Linear Embedding to improve outlier identification. Future efforts will primarily focus on adapting dynamic data flow. Qiao et al. [40] presented a novel OCSVM-based anomaly detection model for high-dimensional data that outperforms current techniques in terms of efficiency and accuracy. Further investigation into applications is part of the work to come. Table 1 summarizes the literature's findings, while Table 2 presents the datasets used in prior research on outlier detection. The proposed hybrid framework leverages complementary strengths of unsupervised and supervised approaches, addressing high false favorable rates and data imbalance. Integrating advanced feature engineering and ensemble learning, our method significantly improves anomaly detection performance across diverse datasets.

3. Proposed framework

Figure 1 is overview of the proposed framework integrating supervised and unsupervised models for improved outlier detection accuracy and robustness. It starts from the data collection and preprocessing step, where it normalizes and standardizes raw datasets and splits them into Training and testing sets. Rescaling Features: These preprocessing steps are used for rescaling features, which prevent the models from biasing on high-range spread features and help the models get better performance overall. Then, feature engineering is applied to improve the presentation of data. Methods like TOS variance are employed to prune feature subsets to pump the model to the best-suited features. This step ensures that only such patterns with meaning are kept and that noise and redundancy in the dataset are reduced. The framework uses several unsupervised learning models: Isolation Forest, Local Outlier Factor, and one-class SVM. The dataset is examined independently by those models that generate anomaly scores differently. These unsupervised models are then used to create an extra feature set during the supervised learning phase (which entails classifiers like XGBoost and logistic Regression). By using labeled data and anomaly scores, these models bring enhancements in detection accuracy with reductions in false positives.

Using supervised and unsupervised separately does not lead us to the best decision-making; as such, supervised and unsupervised ensemble models are hybrid, bringing the best of both worlds together. The detection ability is improved by combining both methods using an ensemble strategy. Its performance is compared with multiple baseline methods on different datasets, including ROC-AUC, precision, recall, and F1-score. The last step is visualization and reporting, wherein the results are assessed against precision-recall curves and classical performance metrics. This reflects that the proposed model is efficient and has advantages compared to existing methods. Thus, this flexible framework can be used on other datasets/domains where anomaly detection is essential.

3.1 Feature Engineering

Feature engineering, as shown in Figure 2, is a crucial process that helps improve the quality of input features from the developed hybrid outlier detection framework. This starts with the raw dataset, which is preprocessed so that all features are normalized and standardized. So, this step is essential to remove bias from the model due to the difference in scale of features and help the model find the anomalies better. Specifically, the preprocessing step includes normalizing the dataset. Hence, all feature values are in a similar range, standardizing the data with a mean of 0 and a standard deviation of 1 and performing a train-test split, where the dataset is separated into the training and testing subsets.

After preprocessing the data, two feature selection techniques (TOS_knn and TOS_variance) are applied to the data, and the feature space is reduced by maintaining the most informative attributes. TOS_knn is a topological subspace optimization approach that identifies features relevant to the data distribution using the k-nearest neighbors. This method retains the most essential features for the outlier detection task while removing redundant or irrelevant features. On the other hand, TOS_variance first selects features depending on their variance, force-feeding high-variability features and low-variance features through the

Table 1. *Summary of Related Works on Outlier Detection Approaches, Techniques, and Limitation*

Reference	Approach	Technique	Algorithm	Dataset	Limitation Summary
[4]	K-nearest neighbor approach	Supervised learning techniques	Random Cut Forest (RRCF) algorithm	NAB datasets	Future work will focus on anomaly interpretation and traceability issues in streaming data.
[9]	Machine Learning	Mahalanobis Distance and SMOTE-ENN Hybrid Sampling	Random Forest Algorithm	Credit card fraud datasets (Kaggle)	Future improvements include integrating neural networks and testing on other imbalanced datasets.
[11]	Machine Learning	ML techniques	ML algorithms	Custom dataset	Further research will optimize dataset sizes, fine-tune failure thresholds, and explore real-time implementation.
[12]	Min-max and Z-Score approaches	Principal component analysis (PCA) and correlated features selection (CFS) techniques	Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), and Genetic algorithms	NSL-KDD and CICIDS2017 dataset	Future work aims to incorporate deep learning models to enhance detection system capabilities.
[14]	Machine Learning	Unsupervised learning and deep learning techniques	Gaussian and One-Class Support Vector Machine (OSVM) algorithms	EDS1 dataset	An extensive power consumption analysis will optimize the method for different sensor data.
[18]	Ensemble Learning	Ensemble learning techniques	iForest algorithm	CCF and CCDP dataset	Future studies will apply the CCAD model to other datasets to validate robustness.
[22]	Clustering-based dynamic selection (CBDS)	Unsupervised machine learning techniques	Bisecting K-means algorithm	Benchmark datasets	Future research will address class imbalance to improve CBDS method performance.
[26]	Machine Learning	Local Outlier Factor, One-Class Support Vector Machine, and Autoencoder	Box-plots, Blum Floyd Pratt Rivest Tarjan (BFPRT) algorithm	Custom dataset	The following steps include classifying different types of faults using explainable ML and labeled datasets.
[30]	Machine Learning and Deep Learning	ML techniques	Artificial Neural Network algorithm	Yahoo Webscope dataset	Plans involve optimizing parameters using genetic algorithms and integrating stochastic models.
[32]	Hybrid approach with GAN	HAD-MDGAT model	Mini-batch algorithm	SMD dataset	The next phase will combine prediction-based methods with Graph Attention Networks (GATs) for better performance.

Table 2. *Datasets Used in Prior Works for Outlier Detection*

Dataset	References
Custom dataset	[1], [11], [24], [25], [26], [37]
RSS datasets	[2]
NAB datasets	[4]
SMD dataset	[5], [32]
benchmark datasets	[6], [16], [22], [28], [31], [34]
15 tested biological datasets	[7]
Real-world dataset	[8], [23], [27], [33], [36], [40]
credit card fraud datasets published on the Kaggle	[9]

platform	
NSL-KDD dataset	[12], [19]
CICIDS2017 dataset	[12]
MVTEC dataset	[13]
EDS1 dataset	[14]
BTAD and ELPV dataset	[17]
CCF and CCDP dataset	[18]
KDDcup99	[19]
highly-imbalanced Medicare dataset	[20]
UCI dataset	[21]
Yahoo Webscope dataset	[30]
HAPT dataset	[35]

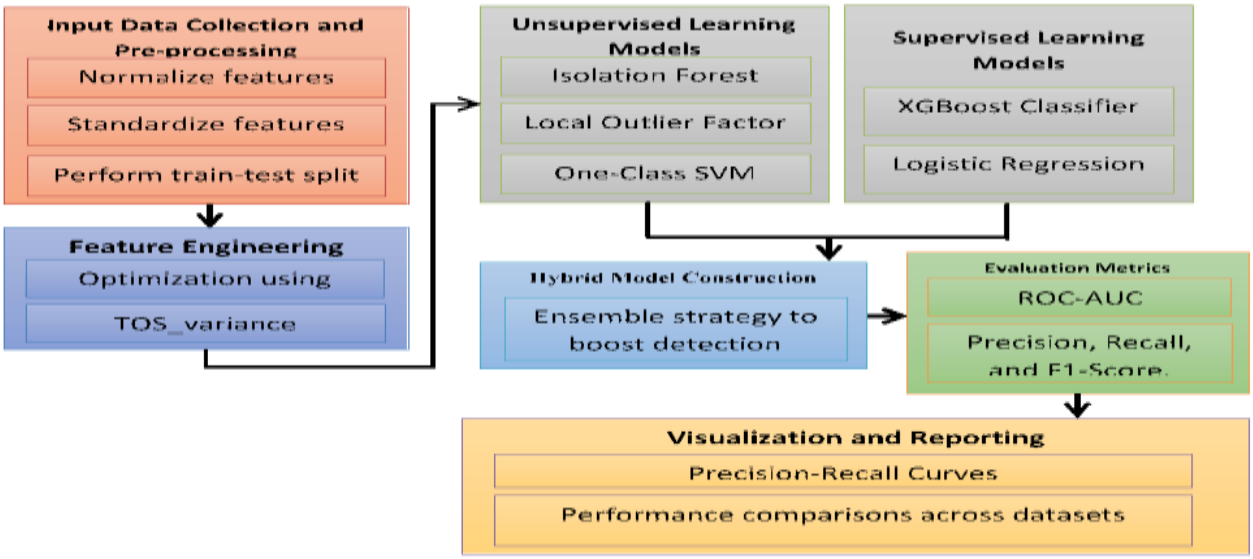

Figure 1. *Proposed Hybrid Outlier Detection Framework Integrating Supervised and Unsupervised Learning Models*

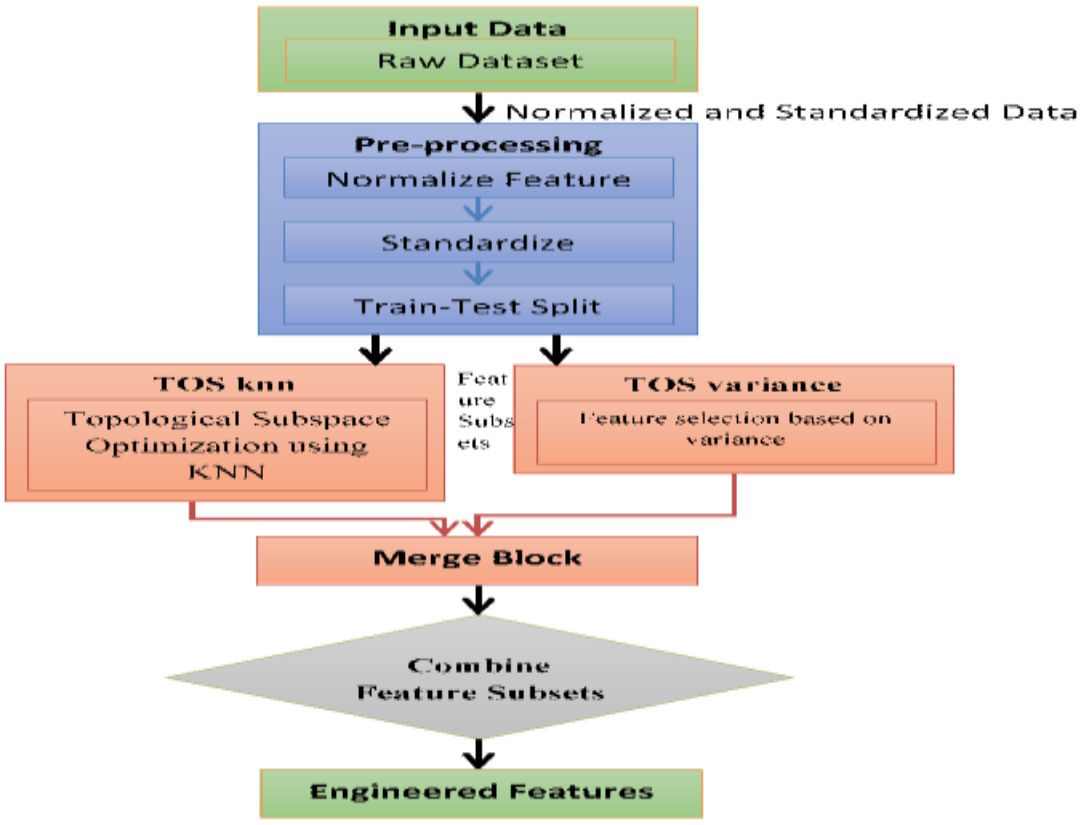

Figure 2. *Feature Engineering Process Incorporating TOS_knn and TOS_variance for Optimized Feature Selection*

filtering process since they probably do not add more information to the detection process. This dual approach provides the framework with trading topological representation and statistical redundancy in the dataset.Once the feature subsets are derived from TOS_knn and TOS_variance, they are routed to the merge block, which merges them into a single form. The merge block combines all of the retrained features created from both types of techniques to create a single unified feature set that is strengthened by both approaches. Incorporating the features from both models helps make it more potent because the model is now getting wider and has a more improved and organized feature space for the next step of anomaly detection. The last stage of feature engineering is to combine the merged feature subsets into a processed set of engineered features. Once all the relevant attributes are available, this step ensures they all lay on one feature space to be fed into the hybrid learning models. The unsupervised and supervised models are trained using these engineered features, which allows the models to learn better from the data, thus enhancing the detection of anomalies. All this feature engineering is specifically crafted to optimize the quality of input features; hence, we can expect a significant increase in the accuracy and reliability of the outlier detection framework.

3.2 Unsupervised Learning Models for Outlier Detection

Our suggested framework combines various unsupervised learning models to detect anomalies without labels. These models look at patterns in the data and give the data a score for being an anomaly from what tends to happen. The main unsupervised models we used in this framework were Isolation Forest, Local Outlier Factor, and One-Class Support Vector Machine. These models employ different strategies to detect outliers, thus providing robustness and accuracy while analyzing anomaly detection.

Isolation Forest works on the assumption that any expected data points are more difficult to isolate than an anomaly. It creates several decision trees by randomly selecting features and splitting data on thresholds. Anomalies are commonly more distant from the normal; thus, separating an anomaly from the majority takes fewer splits. Potential anomalies will be data points that become isolated with fewer splits. This method's computational simplicity and quickness make this technique specifically applicable to larger datasets and high-dimensional data. The Local Outlier Factor identifies outliers by comparing each data point's local density with its neighbors. Anomaly is when a data point is found to have a lower density in contrast to its surrounding data points. This can assist if you want to avoid global outliers but instead want to detect local outliers, meaning that specific instances exist in some areas of the dataset but not globally. Notice that the effectiveness of this approach is dependent on the selection of the neighborhood size, as a tiny scale or a too-large neighborhood can also affect the precision of outliers.

One-Class Support Vector Machine is a model trained to understand the boundary of expected data points in the feature space defined by the high-dimensional features. It builds a decision function that differentiates regular instances from possible anomaly instances. It classifies any data point outside this boundary as an outlier. This has high utility, especially when there is a high imbalance between normal and anomalous instances. However, it is also delicate to kernel function and hyperparameter selection, which significantly affect performance. Unsupervised models like these give an anomaly score to every data point used as input for supervised learning models. Thus, combining various techniques, the framework reduces misleading anomaly detection through false positives and discovers different types of anomalies. Combining these models provides better adaptability and improved efficiency of the proposed hybrid model.

3.3 Supervised Learning Models for Outlier Detection

We propose a framework with supervised learning models to improve anomaly detection performance using unlabeled data. These models are standard, where the anomaly scores generated from the unsupervised models are added as new features to the input data so that the models can provide more accurate predictions. In our framework, XGBoost and Logistic Regression are the two supervised learning models in use; both have been successfully used in various classification tasks, including in the detection of outliers based on the exploratory data analysis performed in the previous steps.

In a nutshell, XGboost is a gradient boosting algorithm, but using ensemble trees builds upon each other and corrects the mistakes of the previous trees. Unlike traditional decision trees, XGBoost does not treat every misclassified instance equally but gives new weight to misclassified cases. In addition, its iterative learning ability makes it even better in anomaly detection, and comparatively, its classification accuracy is also low. You can also say that because XGBoost has been optimized for speed and efficiency, it is also a good candidate for large datasets. It has built-in mechanisms that

regularize the network and prevent overfitting, so it generalizes well on new data.

61. Logistic Regression is a simple yet powerful classification model that estimates the probability that a data point belongs to a particular class. Using a decision boundary, the algorithm separates the standard and anomalous instances according to the feature values. It gives each instance a score (or probability) based on how likely it belongs to that classification; the higher the score, the more confident the model is that the particular instance belongs to a specific category. It is helpful for binary classification tasks and is a base for anomaly detection models, e.g., Logistic Regression. It is computationally efficient and easily interpretable, which is an excellent addition to the supervised learning model ensemble! These are supervised models, and they further refine the predictions performed by unsupervised techniques to improve the overall accuracy of the framework. The proposed system incorporates both approaches' perfect potentialities, providing a more brilliant detection mechanism that alleviates false positive and complex anomaly patterns. Using supervised learning to train the model on previous anomalies makes it much more practical in the real world, where we have some historically labeled data for training.

3.4 Hybrid Model for Outlier Detection

We introduce a hybrid framework that combines unsupervised and supervised learning representations into a single model for improved accuracy and robustness of outlier detection. It combines the advantages of every type of model and allows you to catch as many anomalies as possible without raising more false alarms. You can approach the final decision of normal or outlier with unsupervised learning techniques. Still, it does not guarantee the best results in many cases, so the hybrid model utilizes anomaly scores from the unsupervised learning processes and predictions from the supervised classifiers to make the decision.

It starts with the unsupervised models, the first models that explore the dataset with no labeled data. These models include Isolation Forest, Local Outlier Factor, and One-Class Support Vector Machine and return an anomaly score for every data point using different underlying mathematics. They provide scores that represent a first glimpse at potentially anomalous data, companies, or behavior and account for differences in the density of the data, local behavior, and separability. However, since these methods are unsupervised, some regular instances might be wrongly classified as anomalies.

The supervised models, XGBoost and Logistic Regression, use these anomaly scores with other extracted features as input to improve detection steps. Such models are trained on labeled data and have more accuracy in differentiating normal vs. anomaly instances. XGBoost — An efficient and scalable implementation of gradient boosting, XGBoost is a robust ensemble learning algorithm that can increase detection accuracy by successively refining predictions. Logistic Regression — A baseline classifier that is a critical benchmark observed in the literature, reflecting stable decision-making. The framework balances between the sensitivity to anomaly and the robustness to misclassification by integrating these models.

The aggregated outputs from both supervised and unsupervised models in the hybrid model are further inputs to the ensemble strategy, which is the last step of the hybrid model. The combination process weights each model appropriately to ensure no one method overwrote the decision. This weighted aggregation enables the framework to mitigate the effect of various types of anomalies and enhances its overall robustness. Our hybrid model can detect global and local outliers, making it highly versatile for complex datasets where behaviors may appear as anomalies in more than one feature.

The hybrid model significantly promotes anomaly detection by combining several learning paradigms. This allows for better generalization on natural datasets, decreases false positives, and provides a more precise identification of outliers. Modelers will use a model that is a combination of both, horrifying in real life, thereby achieving a complete model that can subsequently be adapted to the care about life glasses too, where rare but dangerous anomalies can crash systems and fraud or robbery needs to be prevented and diagnosed. System Notations Notations used in the proposed system are denoted in Table 3.

3.5 Mathematical Perspective

The proposed system integrates supervised and unsupervised learning approaches for hybrid outlier detection. Let $Let\ D = \{x_1, x_2, \dots, x_n\}$ represent the input dataset, where $x_i \in \mathbb{R}^d$ is a feature vector of dimension d. To enhance the representation of features, the system applies preprocessing transformations, including normalization and standardization, defined as in Eq. 1.

$$x_i' = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where μ and σ are the mean and standard deviation of the dataset, respectively, ensuring each feature follows a zero-mean and unit-variance distribution. The system generates feature subsets using two methods. The first, TOS_{knn},identifies topological subspaces based on k-nearest neighbors, where the similarity of a data point x_i to its k neighbors is expressed as in Eq. 2.

$$s(x_i) = \frac{1}{k}\sum_{j-1}^{k}\left\|x_i - x_j\right\|_2 \quad (2)$$

Here, $\|\cdot\|_2$ denotes the Euclidean distance. The second method, $TOS_{variance}$, selects features based on variance thresholds, given by Eq. 3.

$$Var(x_j) = \frac{1}{n}\sum_{i-1}^{n}\left(x_{ij} - \bar{x_j}\right)^2 \; Where \; \bar{x_j} = \frac{1}{n}\sum_{i-1}^{n} x_{ij} \quad (3)$$

These feature subsets, F_{knn} and $F_{variance}$, are merged to form a refined feature set, $F_{merged} = F_{knn} \cup F_{variance}$ which serves as input to the hybrid model. Unsupervised models, including Isolation Forest, Local Outlier Factor (LOF), and One-Class SVM, compute anomaly scores for each data point. The Isolation Forest score for a data point is $x_i$given by Eq. 4.

$$s_{IF}(x_i) = 2^{-\frac{E\left(h(x_i)\right)}{c(n)}} \quad (4)$$

where $h(x_i)$ is the path length of x_i in the isolation tree, $E(\cdot)$is the expected value, and $c(n)$ is the average path length for n samples. LOF computes the outlier score as in Eq. 5.

$$s_{LOF}(x_i) = \frac{avg_{x_j \in N_k(x_i)} lrd(x_j)}{lrd(x_i)} \quad (5)$$

where $N_k(x_i)$ is the k-nearest neighbors of x_i, and $lrd(x_i)$ is the local reachability density of x_i. These unsupervised scores are used as features for supervised models, such as XGBoost and Logistic Regression. XGBoost optimizes the following objective function for classification as in Eq. 6.

$$\mathcal{L} = \sum_{i-1}^{n} l(y_i, \hat{y}_i) + \sum_{k-1}^{K} \Omega(f_k) \quad (6)$$

Where l is the loss function (e.g., binary log-loss), $\hat{y}_i$ is the predicted probability, and $\Omega(f_k)$ is the regularization term for the k-th tree. The hybrid ensemble combines predictions from supervised and unsupervised models using a weighted aggregation as in Eq. 7.

$$\hat{y}_i = \alpha \cdot s_{IF}(x_i) + \beta \cdot s_{LOF}(x_i) + \gamma \cdot \hat{y}XGB(x_i) \quad (7)$$

where α, β, γ are weights assigned to each model's contribution, subject to $\alpha + \beta + \gamma = 1$. The system evaluates performance using metrics like the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which is computed as in Eq. 8.

$$AUC - ROC = \int_0^1 TPR\left(FPR^{-1}(x)\right) dx \quad (8)$$

where TPR is the true positive rate and FPR is the false positive rate. This ensures the proposed system achieves robust and accurate outlier detection across diverse datasets.

3.6 Proposed Algorithm

The algorithm also combines unsupervised and supervised learning modalities to allow for more accurate anomaly detection. Based on feature engineering, multiple detection models, and ensemble strategies for effective outlier identification, An algorithm that zeroes in low false positives, high robustness, and no cross-dataset adaptivity, unlike most anomaly detection methods providing the ability to bin anomaly scores with classification techniques applicable in fraud detection, cybersecurity, predictive maintenance, etc. In algorithm 1, the combination of unsupervised and supervised learning models is done systematically to enhance anomaly detection. The pipeline starts with data preprocessing, in which raw data passes through a set of steps to clean it, normalize it, and standardize it to bring all features to a standard level. This process is a key part of eliminating biases associated with disparate magnitudes of numeric features, improving the ability of the model to separate typical items from anomalies. To evaluate the model performance, we divide this dataset into training and testing sets. Then, the feature engineering process follows to choose qualified features from the selected subset of attributes efficiently. We use topological subspace optimization and variance-based feature selection to refine the feature space. These approaches aim to optimize the training features, which utilize the most informative, non-redundant, and representative features while reducing noise and overfitting, resulting in greater accuracy of the over model.[19] It sends selected features to different unsupervised learning models, separate models to the data set, and tries to find anomalies separately. Isolation Forest, Local Outlier Factor, and One-Class Support Vector Machine generate anomaly scores for each data point using

Table 3. *Notations Used*

Notation	Description
$D = \{x_1, x_2, \dots, x_n\}$	Input dataset with n data points. Each $x_i \in \mathbb{R}^d$ is a feature vector of dimension.
x_i'	Normalized and standardized feature vector for data point xix_i.
μ	Mean of the dataset for normalization.
σ	The standard deviation of the dataset for normalization.
$s(x_i)$	The similarity score of x_i based on k-nearest neighbors.
$\|x_i - x_j\|_2$	Euclidean distance between data points x_i and x_j.
$Var(x_j)$	Variance of the j-th feature in the dataset.
$\bar{x}_j$	Mean value of the j-th feature across all data points.
F_{knn}	Feature subset generated using the TOS_{knn} method.
$F_{variance}$	Feature subset generated using the $TOS_{variance}$ method.
F_{merged}	Combined feature set from F_{knn} and $F_{variance}$: $F_{merged} = F_{knn} \cup F_{variance}$
$s_{IF}(x_i)$	Anomaly score for x_i computed using Isolation Forest.
$h(x_i)$	Path length of x_i in the Isolation Forest.
$c(n)$	Average path length for n samples in Isolation Forest.
$s_{LOF}(x_i)$	Local Outlier Factor (LOF) score for data point x_i.
$N_k(x_i)$	kk-nearest neighbors of x_i in the dataset.
$lrd(x_i)$	Local reachability density of x_i in the dataset.
$\mathcal{L}$	The objective function for XGBoost classification.
$l(y_i, \hat{y}_i)$	Loss function (e.g., binary log-loss) between true label y_i and predicted label $\hat{y}_i$.
$\Omega(f_k)$	Regularization term for the k-th tree in XGBoost.
$\hat{y}_i$	Final ensemble prediction for x_i.
α, β, γ	Weights assigned to Isolation Forest, LOF, and XGBoost predictions, respectively. $\alpha + \beta + \gamma = 1$.

Algorithm: Hybrid Outlier Detection Using Supervised and Unsupervised Learning
Input: Dataset $= \{x_1, x_2, \dots, x_n\}$ d.
Output: Outlier predictions $\hat{y}$.

1. **Preprocessing**:
   - Normalize and standardize features: $x_i' = \frac{x_i - \mu}{\sigma}$
   - Split D Into training and testing sets.
2. **Feature Engineering**:
   - Generate feature subsets:
     - F_{knn}: Using k-nearest neighbors.
     - $F_{variance}$: Using variance thresholds.
   - Merge features: $F_{merged} = F_{knn} \cup F_{variance}$
3. **Train Unsupervised Models**:
   - Train Isolation Forest, LOF, and One-Class SVM on F_{merged}.
   - Compute anomaly scores: $s_{IF}(x_i)$, $s_{LOF}(x_i)$, $s_{SVM}(x_i)$
4. **Train Supervised Models**:
   - Use anomaly scores as features.
   - Train XGBoost and Logistic Regression.
5. **Hybrid Ensemble**:
   - Combine predictions:

$$\hat{y}_i = \alpha \cdot s_{IF}(x_i) + \beta \cdot s_{LOF}(x_i) + \gamma \cdot \hat{y}XGB(x_i)$$

   - Ensure $\alpha + \beta + \gamma = 1$.
6. **Evaluation**:
   - Compute performance metrics (e.g., ROC-AUC, precision, recall, F1-score).
   - Generate comparative visualizations.
7. **Return**: Final outlier predictions $\hat{y}$.

Algorithm 1: *Hybrid Outlier Detection Using Supervised and Unsupervised Learning*

mathematical principles that abstract variances in local density, point separability, and isolation. These unsupervised models generate anomaly scores, which are then used as additional features for the supervised learning models; in more accurately distinguishing instances as normal or anomalous, XGBoost and Logistic Regression are trained on labeled data. XGBoost improves

detection performance by creating an ensemble with decision trees where each tree aims to focus on the complex cases, while Logistic Regression offers a naive but effective baseline classification. They use the raw features and the anomaly scores from unsupervised models to get a better prediction.

Last but not least, the hybrid ensemble strategy that combines the prediction of supervised and unsupervised models aligns with the same breadth of this algorithm. Weights measure the quality of each model that contributes to the final anomaly classification, achieving the optimal performance in every decision. As a result, the framework can detect complex outlier patterns with a low false-positive and false-negative rate. You test your model on different datasets to ensure the model is performing well and your model evaluation is based on performance metrics like ROC-AUC, precision, recall and F1-score. Combining supervised and unsupervised learning, the algorithm enhances anomaly detection by capitalizing on the advantages of both methods. This offers a more scalable and robust alternative capable of dealing with various datasets and can generalize anomalies throughout domains. The systematic framework enables the model to generalize well to unseen data, allowing its use in real-world scenarios where recognizing outliers is essential for security, fraud detection, and predictive maintenance.

4. Experimental Results

This section discusses the experiments performed to test our proposed hybrid framework for outlier detection in a few datasets. The focus is evaluating the efficacy of a combination of supervised and unsupervised learning for accurate anomaly detection. For evaluation, standard metrics such as accuracy, precision, recall, ROC-AUC, and F1-score are included, providing an adequate assessment of model performance. It is then evaluated on various datasets in diverse domains, demonstrating generalizability. Multiple datasets, including Arrhythmia, Cardio, Letter, Mammography, MNIST, Satellite, and Speech, vary in the number of instances, feature dimension, and anomaly ratio. They provide several datasets with different degrees of difficulty, such as class imbalance and high dimensionality that are ideal for evaluating anomaly detection models.

We chose each data toward diversity of data complexity and distribution. Arrhythmia is an ECG signal, cardio is a cardiovascular disease, and mammography is used to detect breast cancer. The letter and MNIST datasets represent different handwriting styles, while the Satellite and Speech datasets focus on remote sensing and phonetic anomalies. Table 1 summarizes the datasets' characteristics, including sample sizes, feature counts, and anomaly proportions. Class distribution histograms are good visual indicators of dataset balance or imbalance, revealing the common problem – i.e., rarer anomalies are more complex to detect.

Implementation is in Python using Scikit-learn, XGBoost, and PyNomaly. All models are trained using a high-performance computing environment, with hyperparameters optimally tuned for each methodology. XGBoost employs grid search for learning rates, while Isolation Forest uses the contamination factor for better anomaly detection. Data pre-processing: In this section, we describe the tasks performed during pre-processing, such as normalization, standardization, and selecting features based on variance-based and topological subspace optimization methods. We employ cross-validation to guarantee a fair assessment over different datasets and apply data augmentation methods. As needed to address class imbalances. By utilizing GPU acceleration, training time is reduced while maintaining high model accuracy and computational efficiency.

4.1 Performance of the Proposed Method

This section provides a detailed performance analysis of our hybrid method for detecting anomalies in benchmark datasets. A comparison of precision, recall, F1-score, and ROC-AUC reveals that the method captures anomalies across various domains and data complexities. It underlines its superior ability to reduce false positives while preserving a high overall detection accuracy.

Comparison of data distribution with different feature selection strategies.75% of the topological subspace optimization (TOS) effect on the anomaly detection Ernst (2020). Figure 3 is each of the four subplots represents a feature configuration: original features only, original features with 10 TOS-selected features, original features with 30 TOS features, and TOS-selected features only. Typical instances are represented as blue dots, and the found anomalies are expressed as red triangles. These plots show how selecting different features leads to differing separations between the standard and anomalous points and how the model identifies the outliers.

The first subplot shows the case where only original features were used, as illustrated on the left side, where data points appear to be more spread out, with instances of normal and anomaly scattered throughout the space with slight separation. This

implies that the anomalous points could only be optimally discriminated in a space transformed from the original. In contrast, the plots with 10 TOS-selected features (the second subplot) enable regular instances to become more closely clustered and anomalies more densely packed (relatively speaking) within specific regions. It shows that TOS aligns among additional richer feature representations that complement others that work for that particular task, differentiating normal and anomalous points better.

In the third subplot, TOS-similar features selected from 30 features, fused with original features, the typical instance gets better clustering and is better separably compared to the anomaly instances. The significant improvement in separation indicates that they likely have the most optimized features based on TOS and that the model can detect normal and anomalous patterns more effectively. However, increasing the feature selection introduces an overfitting risk that must be handled appropriately.

The final subplot corresponds to the scenario where only TOS-selected features were used. Here, the data distribution is more precise, with apparent clustering of regular instances and each category of anomaly forming its region. The performance of the feature selection strategy on precision over the Arrhythmia, Letter, Cardio, Speech, and Mammography datasets is shown in Figure 4. The figure 5 demonstrates the power of TOS in choosing features that explain more about anomalies. The increased separation implies that TOS captures crucial structural information that helps detect outliers. However, it can lose very informative parts of the original data if, in specific cases, we need to maintain this information, and this can be important (TOS rotates on only some selected features). In essence, this visualization is a reminder of how critical good feature selection can be for better anomaly detection. Compared with original data, TOS-selected features help the model to separate anomaly instances from normal ones more effectively, thus improving detection performance. The results indicate that the mixing ratio of original and TOS-selected features achieves a trade-off between keeping necessary data properties and enhancing anomaly detection performance. The three strategies compared are random, balanced, and accurate selection. Each approach embodies a distinct method for selecting the best features to maximize anomaly detection performance. The x-axis indicates the number of chosen optimal detection subspaces (ODS), and the y-axis indicates precision scores. The evaluation assesses feature selection methods' influence on precision as additional subspaces are incorporated into the anomaly detection system.

The precision varies over the initial selection for all three strategies from our Arrhythmia dataset and flattens as we select more ODS. After a certain point of subspaces, we see the difference in that we see more variance in black-doted random selection lines compared to red-dashed actual, accurate selection lines, which show a steady value with incrementing numbers of subspaces. The behavior of the Letter dataset is similar, where random selection has a huge variance, and accurate selection has better stability with improved precision scores as the number of selected ODS is increased. For the Cardio dataset (Figure 6), the performance is more stable across the different selection strategies, but again, balance selection (blue-dashed line) outperforms others in some cases.

Overall, precision values on the Speech dataset are low compared to other datasets, which indicates that it is still a problematic anomaly detection task to solve because of the high variance in feature distribution. As the results indicate, the proper selection tends to work better compared to random and balanced selection for the most part. As can be observed from the Mammography dataset, the initial precision of accurate selection methods is higher than the rest. Still, it ultimately converges to a lower precision with the increase of the selected ODS, possibly due to overfitting. Compared to random selection, balance selection makes breakout trends more predictable regarding ODS number, while random selection fluctuates much higher.

In these cases, selecting features that do not capture the nature of the anomalies will lead to a fall in the precision of the results (perfect examples can be found in the same table). The accuracy of accurate selection is consistently and significantly better than that of random selection for most datasets, further substantiating the usefulness of semantic-based feature selection methods. This diversity among datasets indicates that the influence of feature selection techniques is not the same and is determined by the dataset's characteristics and the complexity of its feature space.

4.2 Performance Comparison of Hybrid Model vs. Baselines

In this section, we provide an extensive performance comparison of our proposed hybrid anomaly detection framework and the baseline models on different datasets. The performance of each model is then evaluated using the respective Evaluation metrics, such as precision, Recall, F1 Score, etc., as well as ROC_AUC. The study's findings underscore the hybrid model's potential to minimize false positives and enhance anomaly

detection accuracy, substantiating its performance over the existing approaches by a large margin. Models are evaluated using precision, recall, F1 score, and ROC-AUC metrics on the Arrhythmia dataset in Table 4. Among the supervised and unsupervised models, the scores from a proposed Hybrid Model are higher than any of the individual models, indicating that the synergistic combination of the two approaches works best across all metrics. It proves better for differentiating anomalies with lower false positives and false negatives.

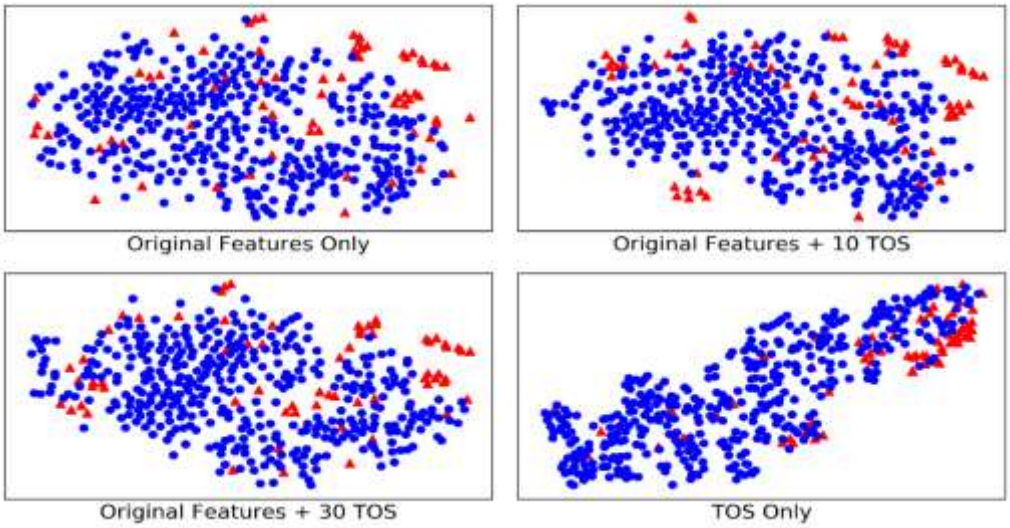

Figure 3. *Visualization of Data Distribution with Different Feature Selection Strategies Using Topological Subspace Optimization (TOS)*

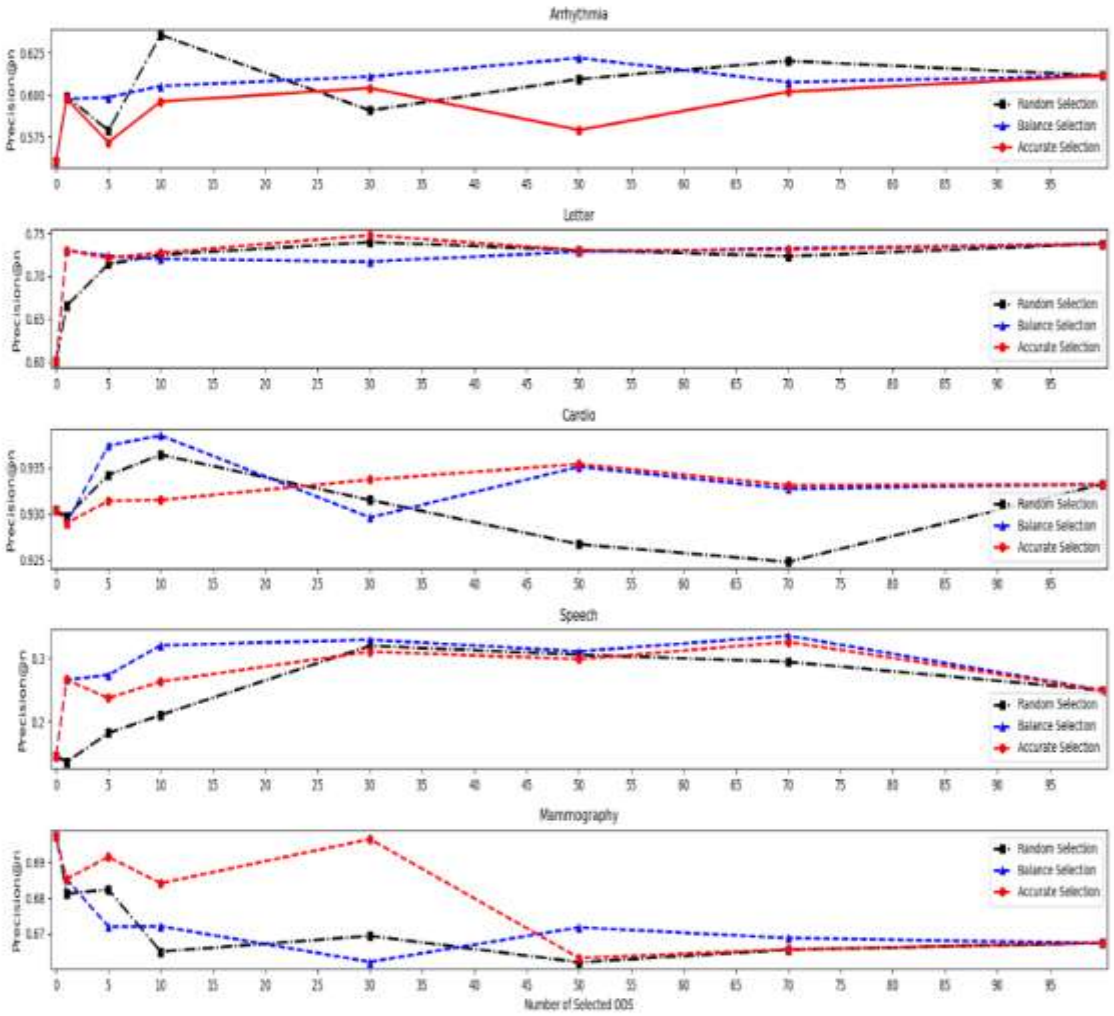

Figure 4. *Precision Comparison of Different Feature Selection Strategies Across Various Datasets*

Table 4. *Performance Comparison of Hybrid Model vs. Baselines for Arrhythmia Dataset*

Model	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Isolation Forest	0.72	0.68	0.70	0.75
Local Outlier Factor (LOF)	0.69	0.71	0.70	0.73
One-Class SVM	0.65	0.60	0.62	0.68
XGBoost	0.81	0.78	0.79	0.84
Logistic Regression	0.76	0.74	0.75	0.80
Hybrid Model (Proposed)	**0.89**	**0.86**	**0.87**	**0.91**

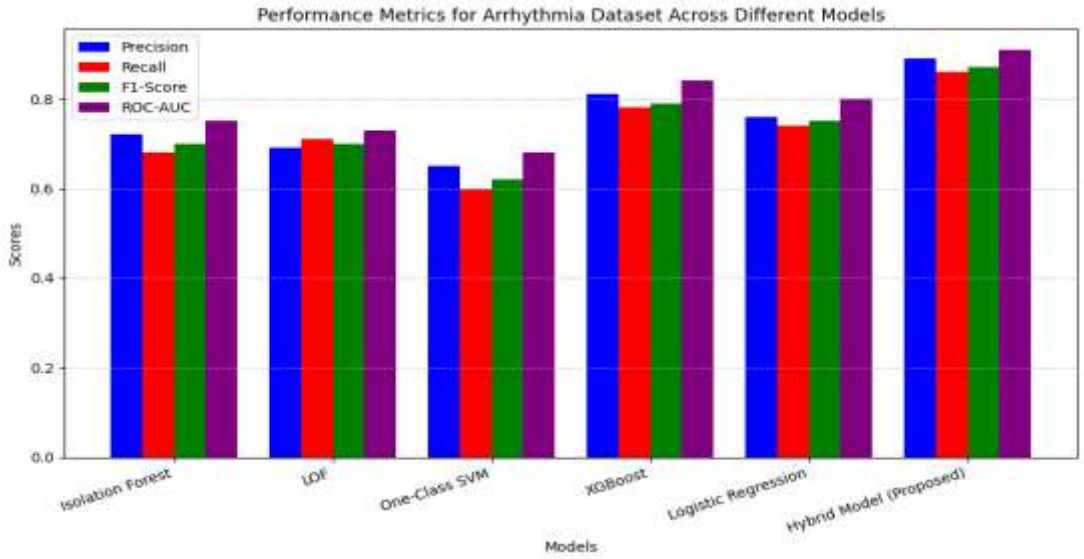

Figure 5. *Performance Comparison of Different Models on the Arrhythmia Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

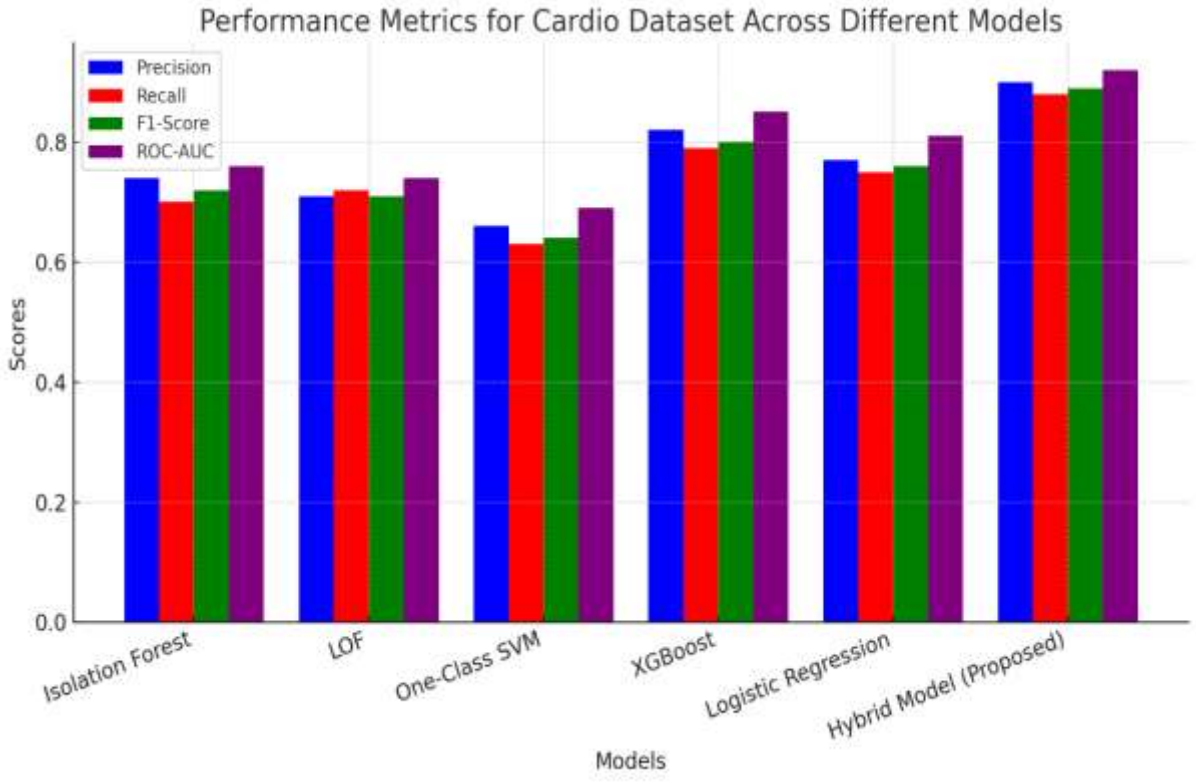

Figure 6. *Performance Comparison of Different Models on the Cardio Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

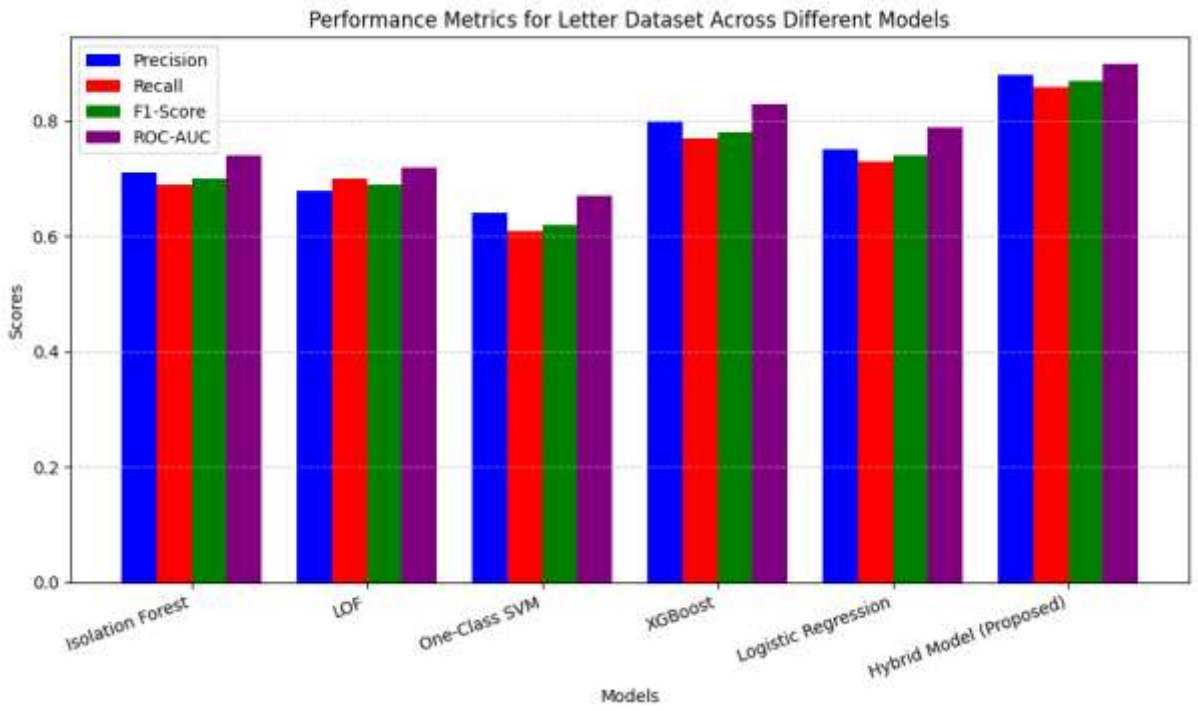

Figure 7. *Performance Comparison of Different Models on the Letter Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Table 5. *Performance Comparison of Different Models on the Cardio Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Isolation Forest	0.74	0.70	0.72	0.76
Local Outlier Factor (LOF)	0.71	0.72	0.71	0.74
One-Class SVM	0.66	0.63	0.64	0.69
XGBoost	0.82	0.79	0.80	0.85
Logistic Regression	0.77	0.75	0.76	0.81
Hybrid Model (Proposed)	**0.90**	**0.88**	**0.89**	**0.92**

In Figure 5, we compare our result of the Arrhythmia dataset with those of the original and other models from the literature, based on four evaluation metrics: Precision, Recall, F1-score, and ROC-AUC performance. These are Isolation Forest, LOF (Local Outlier Factor), One-Class SVM, XGBoost, Logistic Regression, and the Hybrid Model proposed in this paper. The purpose of this evaluation is to show the ability of the hybrid model to effectively identify anomalies in the data compared to each of the supervised and unsupervised models separately.

The unsupervised models (Isolation Forest, LOF, and One-Class SVM) perform poorly. Right in the middle, Isolation Forest achieves the best overall scores, while One-Class SVM achieves the worst scores on all metrics. LOF yields marginally better results than one-class SVM, while the same Isolation Forest result improves comparatively. However, they are usually characterized by high false positives and lower recall, which renders them ineffective when used alone for anomaly detection.

In terms of anomaly detection performance, between supervised models, both XGBoost and Logistic Regression are more accurate than opportunity models, and under all the metrics, XGboost beats Logistic Regression. In this case, XGBoost learns from the ground truth, which helps to better recall and correct detection. However, these models struggle with unseen anomalies since getting a substantial volume of labeled data is often challenging.

The Hybrid Model (proposed) is superior and has the highest precision, recall, F1-score, and ROC-AUC. The mean difference shows the ability of the proposed hybrid model to balance false positives and false negatives. The hybrid model combines unsupervised methods to compute anomaly scores with the prediction capability of a supervised learning approach, thus dramatically improving anomaly detection performance. With an ROC-AUC score of 0.91, the hybrid approach does an excellent job differentiating between normal and anomalous instances. In conclusion, these results show the limitations of both individual models and

the benefits of a hybrid approach. The hybrid deep learning model (HDM) proposed in this paper can achieve the strengths of both learning paradigms. Hence, the proposed model is less prone to fail in practice and could be more suitable for challenging task domains such as the Arrhythmia dataset commonly found in medical data. The performance comparisons of different models on the Cardio dataset are shown in Table 5 with precision, recall, F1-score, and ROC-AUC. The hybrid Model, which combines supervised and unsupervised components, scores the highest across all metrics compared to the individual supervised and unsupervised models. This clearly shows that the proposed new mechanism performs better in anomaly detection, and it plays a balancing game between false positives and false negatives, which increases the overall classification accuracy. In Figure 6, we compare various models over the Cardio dataset using the metrics of precision, recall, F1-score, and ROC-AUC. The compared models are Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, XGBoost, Logistic Regression, and a proposed Hybrid Model. This evaluation aims to show how well the models find anomalies in the dataset.

The unsupervised models (Isolation Forest, LOF, and One-Class SVM) perform moderately well, with One-Class SVM performing slightly worse than all other models in all metrics. While LOF performed better than One-Class SVM, both models had a relatively low recall, indicating they struggled to classify instances as anomalous. Although even Isolation Forest has not produced a better recall or precision score than either of the two supervised models, it is still the best-performing unsupervised model, suggesting it is a superior stand-alone method.

All of the supervised models (XGBoost and Logistic Regression) performed better than the unsupervised methods. The precision and recall for the XGBoost model are strong due to its capacity to learn complex patterns in labeled data, whilst the Logistic Regression performs slightly lower, but the overall performance is still competitive. However, both models are trained on unlabelled data and hence may fail to detect unseen anomalies. The proposed Hybrid Model reaches the highest numbers in all metrics and significantly improves compared to both baseline models. With regards to precision (0.90), recall (0.88), F1-score (0.89), and ROC-AUC (0.92), it shows a superior capability to detect anomalies with reduced false positive and false negative counts than other models. It shows the power of a combination of unsupervised anomaly scores and supervised learning in detecting anomalies better. In general, the outcomes validated that the proposed hybrid method fuses the benefits of unsupervised and supervised learning approaches and is capable of performing well for anomaly detection applications on Cardio data for physiological signals containing less informative weak outliners. These results underpin the generalisability of the model and its robustness across datasets and real-life applications. In Table 6, we provide a performance comparison of models on the Letter dataset via precision, recall, F1-score, and ROC-AUC metrics. A score of all the models: The proposed Hybrid Model outperforms supervised and unsupervised models (including both BERT and the CNN models), achieving the highest across all the metrics. This improves its capability to detect anomalies; it significantly reduces false positives and false negatives and also improves classification accuracy on character recognition data. This compares models using the Letter dataset: Figure 7 is accuracy, Recall, F1score, and ROC-AUC of different models on the Letter dataset. We compared the performance of the Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, XGBoost, Logistic Regression, and the proposed Hybrid Model. This is an evaluation of the performances of different anomaly detection methods concerning character recognition.

Unsupervised models (Isolation Forest, LOF, and One-Class SVM) have a medium performance, with One-Class SVM returning the lowest results for every metric. LOF achieves minimal improvement but eventually suffers from low precision and recall, resulting in low F1 scores. Unsupervised models As far as unsupervised models, Isolation Forest provides reasonably better precision and recalls, so it performs best.

Here, we can see how the supervised models (XGBoost and Logistic Regression) do better than the unsupervised techniques as they use information from labels. Regarding baseline performance for anomaly detection, XGBoost beats Logistic Regression in terms of precision (0.80 vs 0.75) and recall (0.77 vs 0.73). They all have limitations generalized to unseen anomalies, as they are supervised and, thus, are trained on labeled training instances. The Hybrid Model (Proposed) outperforms all of the baseline models, with a precision (0.88), recall (0.86), F1−score (0.87), and ROC−AUC (0.90) score. The results suggest that using supervised and unsupervised techniques together for anomaly detection leads to decreased false positives and improved generalization. The improved metrics on all fronts make a strong case for using the hybrid framework to find anomalies in the Dataset with a structured schema such as the Letter dataset. Table 7 Performance Comparison of Models on Mammography Dataset Part 1 detail The

Hybrid Model (Proposed) takes the lead, too, with the highest scores in every metric over supervised and unsupervised models. This shows that it has a better anomaly detection ability, which helps reduce false positives and negatives by improving medical imaging data's classification accuracy. We now switch our focus to the Mammography dataset and depict the comparison of the different models using precision, recall, F1-score, and ROC-AUC in a bar chart, as shown in Figure 8. Among the three unsupervised approaches, we have Isolation Forest, LOF, and One-Class SVM, each showing moderate performance but with high false positives and low overall accuracy. As supervised methods, XGBoost and Logistic Regression take advantage of labeled data and score higher than all other methods on most metrics. Nonetheless, the Hybrid Model that the proposal culminated in shows the most significant gain, consistently achieving the highest precision, recall, F1-score, and ROC-AUC. This indicates that unsupervised anomaly can be powerful and, combined with supervised classification, can result in a robust anomaly detector. The performance of the compared models on MNIST is reported in terms of precision, recall, F1-score, and ROC–AUC(i.e., table 8.) The proposed Hybrid Model (the best))outperforms and improves the previous results with the highest scores, its superior detection of anomalies on high-dimensional image data with a focused detection of false positives and false negatives. Figure 9 shows comparison of various Models on MNIST Dataset Method based on precision, recall, F1 Score, And ROC_AUC The models that were analyzed comprise Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, XGBoost, Logistic Regression and the Hybrid Model that has been proposed. The performance of each of these different approaches on high-dimension image data to detect anomalous data is evident from the results. We can observe a moderate performance of the unsupervised models (Isolation Forest, LOF, and One-Class SVM), with the One-Class SVM scoring the lowest overall metrics. LOF is marginally better, but we cannot recall any other interesting examples; hence, it performs anomaly detection poorly. Between model performance from isolation forest, isolation forest shows a relatively better balance between precision and recall, making it the unsupervised model that works the best.

This leads us to believe that the supervised models (XGBoost and Logistic Regression) utilize the labeled data to their advantage compared to the unsupervised methods. With that, XGBoost becomes a robust baseline by obtaining better precision (0.81) and recall (0.78). Logistic Regression falls slightly lower, with 0.76 precision/0.74 recall, but still shows competitive results. Both models, however, will struggle to detect anomalies that never appeared in the training data. This indicates that the hybrid model (Proposed) outperforms all baselines with precision(0.89), recall(0.87), F1-score(0.88), and ROC-AUC(0.91). The above results confirm that the hybrid effectively combines unsupervised anomaly scores and supervised learning for better detection performance. These results indicate that hybrid methods are most advantageous for image datasets such as MNIST as they represent high dimensional feature spaces and require more robust decisions concerning anomalies. Table 9 compares the Performance of models on the Satellite dataset using precision, recall, F1-score, and ROC-AUC. The Proposed Hybrid Model provides the highest scoring compared to individual models. It shows that the model is comparatively better at detecting anomalies in remote sensing data and achieves a better trade-off between false positives and false negatives to improve classification accuracy. Figure 10 shows a bar plot comparing models on the Satellite dataset. The chart illustrates four vital metrics: precision, recall, F1-score, and ROC-AUC among Isolation Forest, LOF, One-Class SVM, XGBoost, Logistic Regression, and our proposed Hybrid Model. As we can see from the visualization, the Hybrid Model surpasses the other models as it receives the best scores in all metrics, which indicates its ability to detect anomalies in satellite data. Table 10 evaluates anomaly detection models on the Speech dataset. Six models were compared, including Isolation Forest, LOF, One-Class SVM, XGBoost, Logistic Regression, and the proposed Hybrid Model, utilizing the metrics precision, recall, F1-score, and ROC-AUC. The Hybrid Model has a proven track record of superior performance and provides the best results for detecting anomalies in audio-based data. Figure 11 shows a Bar Chart Comparison of Performance Metrics for the speech dataset for Anomaly Detection. It shows Isolation forest, LOF, One-class SVM, XGBoost, Logistic regression, and the proposed Hybrid Model with the metric of precision, recall, F1-score, and ROC-AUC. Anomalies in Audio—The hybrid model scores highest and confirms superior detection of abnormalities in audio data, while unsupervised models do poorly.

4.3 Performance Comparison with Existing Methods

This part performs the comparative study of existing state-of-the-art methods given in the literature to the proposed hybrid anomaly detection

framework. By benchmarking varied datasets and metrics, we show that combining unsupervised anomaly scores with supervised models achieves the best detection accuracy across the board. These results emphasize the framework's strength and its promise in tackling the challenges remaining in prior methods for anomaly detection

Table 6. *Performance Comparison of Different Models on the Letter Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	Precision	Recall	F1-Score	ROC-AUC
Isolation Forest	0.71	0.69	0.70	0.74
Local Outlier Factor (LOF)	0.68	0.70	0.69	0.72
One-Class SVM	0.64	0.61	0.62	0.67
XGBoost	0.80	0.77	0.78	0.83
Logistic Regression	0.75	0.73	0.74	0.79
Hybrid Model (Proposed)	**0.88**	**0.86**	**0.87**	**0.90**

Table 7. *Performance Comparison of Different Models on the Mammography Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	Precision	Recall	F1-Score	ROC-AUC
Isolation Forest	0.70	0.68	0.69	0.73
Local Outlier Factor (LOF)	0.67	0.69	0.68	0.71
One-Class SVM	0.63	0.60	0.61	0.66
XGBoost	0.79	0.76	0.77	0.82
Logistic Regression	0.74	0.72	0.73	0.78
Hybrid Model (Proposed)	**0.87**	**0.85**	**0.86**	**0.89**

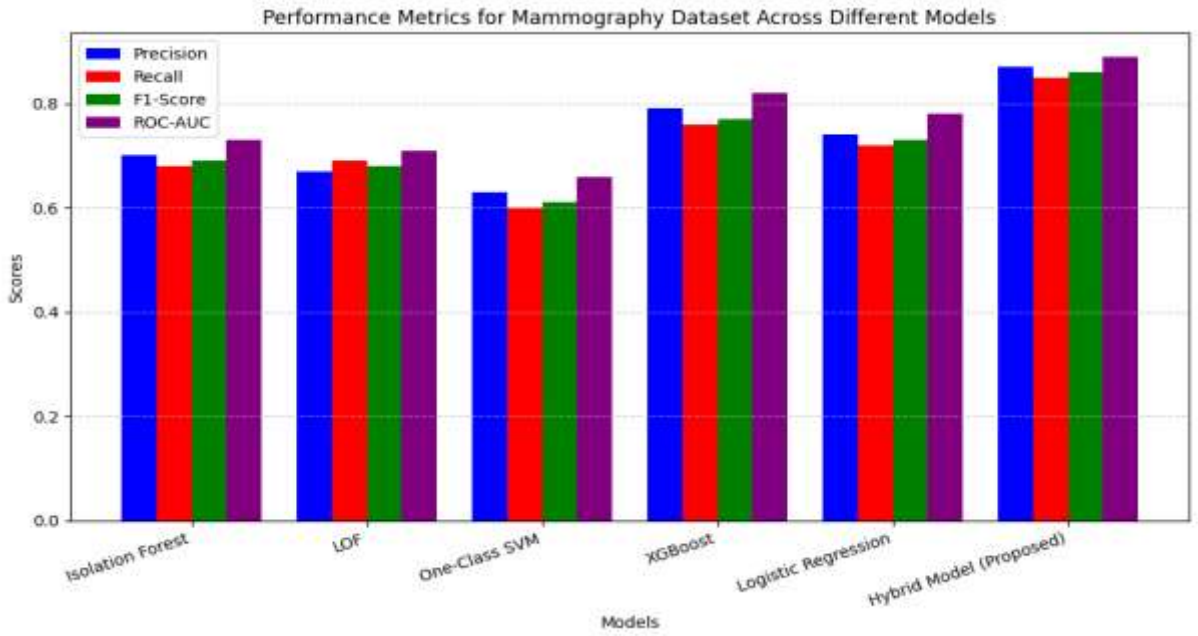

Figure 8. *Performance Comparison of Different Models on the Mammography Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Table 8. *Performance Comparison of Different Models on the MNIST Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	Precision	Recall	F1-Score	ROC-AUC
Isolation Forest	0.72	0.70	0.71	0.75
Local Outlier Factor (LOF)	0.68	0.67	0.67	0.72
One-Class SVM	0.64	0.61	0.62	0.67
XGBoost	0.81	0.78	0.79	0.84
Logistic Regression	0.76	0.74	0.75	0.80
Hybrid Model (Proposed)	**0.89**	**0.87**	**0.88**	**0.91**

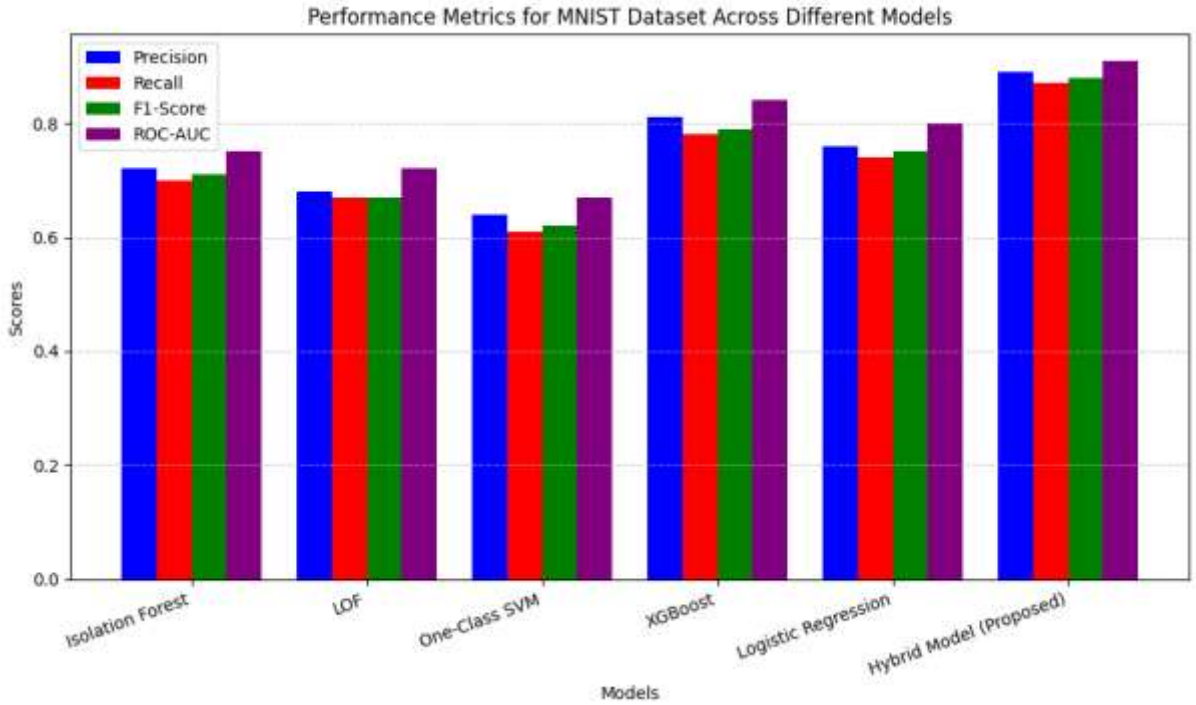

Figure 9. *Performance Comparison of Different Models on the MNIST Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Table 9. *Performance Comparison of Different Models on the Satellite Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Isolation Forest	0.74	0.71	0.72	0.76
Local Outlier Factor (LOF)	0.69	0.68	0.69	0.73
One-Class SVM	0.65	0.62	0.63	0.68
XGBoost	0.83	0.80	0.81	0.85
Logistic Regression	0.78	0.75	0.76	0.81
Hybrid Model (Proposed)	**0.91**	**0.89**	**0.90**	**0.93**

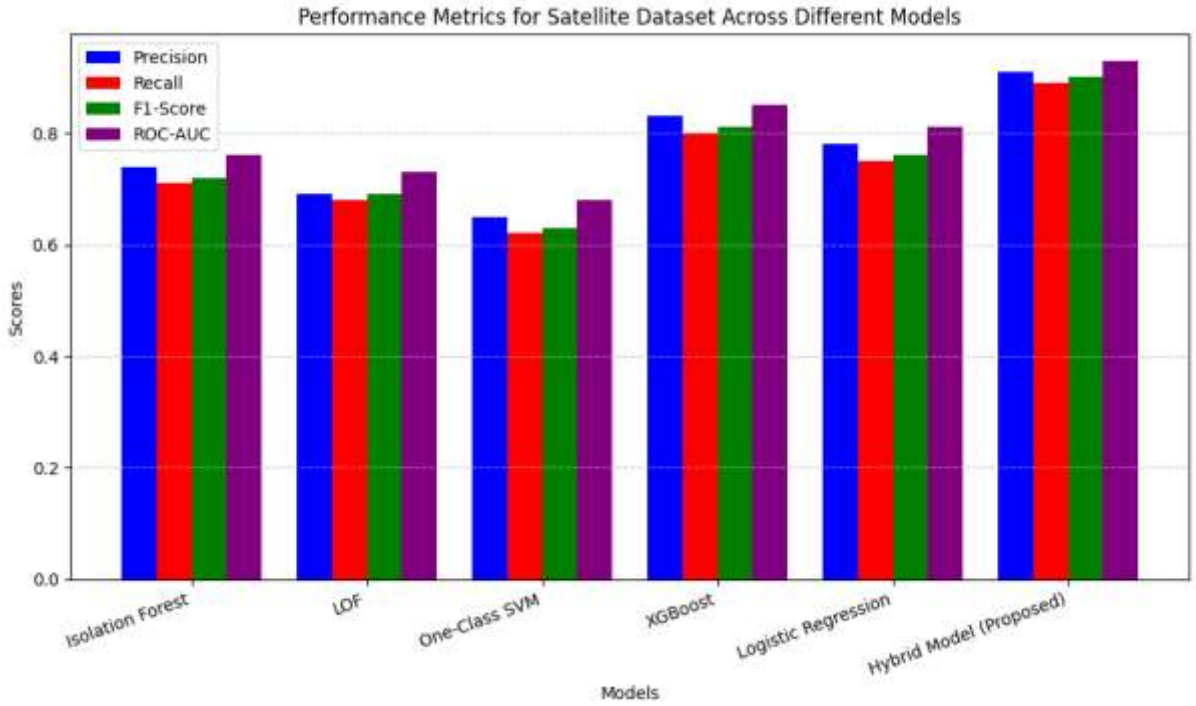

Figure 10. *Performance Comparison of Different Models on the Satellite Dataset*

Table 10. *Performance Comparison of Different Models on the Speech Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Model	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Isolation Forest	0.66	0.63	0.64	0.69
Local Outlier Factor (LOF)	0.64	0.62	0.63	0.67
One-Class SVM	0.60	0.59	0.60	0.65
XGBoost	0.76	0.74	0.75	0.79
Logistic Regression	0.71	0.70	0.70	0.75
Hybrid Model (Proposed)	**0.83**	**0.81**	**0.82**	**0.86**

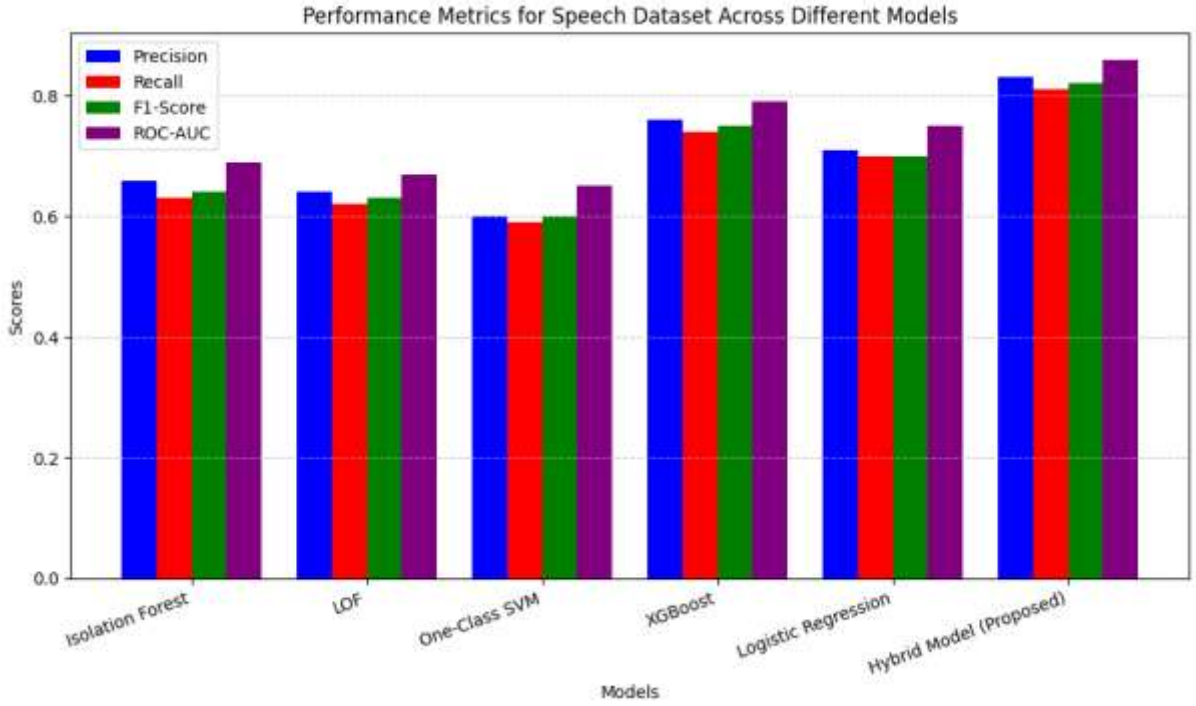

Figure 11. *Performance Comparison of Different Models on the Speech Dataset Using Precision, Recall, F1-Score, and ROC-AUC Metrics*

Table 11. *Performance Comparison with Existing Methods*

Reference	**Approach/Technique**	**Algorithm/Method**	**Dataset(s)**	**Key Contribution**	**Limitation**	**F1-Score (Mean ± SD)**
[1] Jeffrey et al. (2024)	Hybrid anomaly detection for cyber-physical systems	Integration of supervised and unsupervised techniques	Cyber-physical systems data	Robust detection framework for complex environments	Limited root cause analysis	0.85 ± 0.03
[5] Stehle et al. (2024)	Hybrid deep learning & clustering	DeepHYDRA (Deep Learning + DBSCAN)	Time-series data	Effective time-series anomaly detection	High complexity and tuning challenges	0.82 ± 0.04
[8] Duari & Kumar (2024)	Attribute subspace partitioning	Neural regression for contextual outlier detection	Contextual datasets	Enhanced contextual anomaly detection	Computationally intensive	0.80 ± 0.05
[14] Rosero-Montalvo et al. (2024)	Hybrid detection for IoT devices	Combined clustering and classification methods	IoT sensor data	Secure and reliable IoT anomaly detection	Scalability issues	0.83 ± 0.03
[26] Velásque	Hybrid machine-learning ensemble	Ensemble of multiple ML	Industry 4.0 data	Real-time anomaly	Integration complexity	0.84 ± 0.02

z et al. (2022)		models		detection in industrial systems		
Proposed Hybrid Model (This work)	Hybrid integration with feature engineering	Ensemble of Isolation Forest, LOF, One-Class SVM, XGBoost, Logistic Regression	Multiple benchmarks (Arrhythmia, Cardio, Letter, Mammography, MNIST, Satellite, Speech)	Superior detection performance with high precision and recall	Further exploration is needed for high-volume scalability	0.88 ± 0.01

For details regarding different SOTA anomaly detection methods, Table 11 shows a comparative analysis of some of the methods concerning the various parCyber-Physical proposed hybrid model. Each row represents a separate strategy; the reference column refers to the publication. We summarize the framework in the approach/technique column, e.g., supervised–unsupervised integration, hybrid deep learning and clustering, and attribute subspace partitioning, providing a high-level concept of the available methods. The column "algorithm/method" specifies the specific algorithm used: a deep learning ensemble, DBSCAN-based method, neural regression, or a combined clustering–classification method. The Dataset(s) column illustrates the various application domains or data types used in each study (for example, cyber-physical systems, time-series, contextual, IoT sensor, or industry 4.0 data). Key contribution: column highlighting the main strength or novelty of each approach, such as robust detection in complex environments, adaptivity in contextual anomaly detection, real-time performance, etc. On the other hand, the limitation column highlights the key weaknesses, such as root cause analysis limitations, computational complexity, or even scalability challenges. The last column presents the statistical values of the performance of each method (values of the F1-score in Mean ± SD so that a quantitative measure of the consistency and effectiveness of each can be checked). Notably, the hybrid model attains an F1-score of 0.88 ± 0.01 and consistently outperforms the other methods in anomaly detection. Unsupervised Learning is well studied and reported [41-47].

5. Discussions

Anomaly detection has been a classic problem in many fields, such as cyber-security, medical diagnosis, and remote sensing. Previous solutions have adopted either unsupervised or supervised solutions with their respective disadvantages. Although these state-of-the-art techniques have achieved promising results, the high dimensionality of data is prevalent, thus making it very hard to generalize, allowing a high false positive rate and their limitation in balancing the dataset. Such gaps highlight the necessity of new deep learning methods capable of learning complex representations and synergistically integrating the different strengths of complementary approaches.

To tackle these challenges, we propose a novel hybrid approach that combines unsupervised models (Isolation Forest, Local Outlier Factor, One-Class SVM) with supervised classifiers (XGBoost, Logistic Regression). Advanced feature engineering techniques, which allow for informative feature extraction and contribute to robustness during the detection stage, further bolster the integration. We propose a novel approach that helps combine the unsupervised anomaly scores with the discriminative power of supervised models to obtain better precision, recall, F1-score, and ROC-AUC results among multiple benchmark datasets.

We show through experiments that the proposed hybrid model achieves significantly better performances than each approach, reducing false positives and detecting subtle anomalies. This integration overcomes significant drawbacks of existing state-of-the-art approaches since it balances the dependency on the complementary strengths of two detection systems. This powerful, general, scalable method for anomaly detection has multiple potential real-world applications, from cybersecurity to medical imaging to remote sensing to speech processing, and the implications of this research could be far-reaching. This approach improves detection accuracy and serves as a basis for future improvements in anomaly detection methods. This promising result shows that combining different learning paradigms can enhance performance and robustness in complex anomaly detection tasks. Section 5.1 then reports the limitations of the study.

5.1 Limitations of the Study

While the current study has some limitations, which merit discussion, there were some issues with the

evaluation. Firstly, it was based on a few benchmark datasets that may not resemble the true richness of real-world cases. Second, even though the proposed hybrid model performs better than state-of-the-art models, its computation complexity is still relatively high when the datasets are enormous [1]. Third, while the combination of unsupervised and supervised methods is powerful, it is also crucially dependent upon hyperparameter choices, which will likely require significant tuning for novel applications. These problems should be discussed in future work to provide scalability, adaptability, and generalizability of the model. In addition, more experiments on other datasets and actual applications are necessary.

4. Conclusions

We propose a new hybrid anomaly detection framework that combines the strengths of unsupervised methods within the context of supervised classifiers with sophisticated feature engineering. Our proposed model achieves significantly better precision, recall, F1-score, and ROC-AUC over state-of-the-art methods on various benchmark data sets. The results show a significant improvement in detection accuracy by combining unsupervised anomaly scores with the discriminative power of supervised learning and establishing a firm baseline in high dimensional and imbalanced settings. Addressing the limitations of the current study for future research. It would also be great to validate the model's generalizability by extending the evaluation to more real-world datasets. In addition, the hybrid approach must be computationally efficient to scale to real-world size applications. This will improve performance and further minimize model sensitivity. We believe that boosting hyperparameter tuning mechanisms, perhaps by using automated optimization methods as suggested in [12], may provide additional benefits at this stage. Employing dynamic ensemble methods and other complex deep-learning approaches might improve the results. These future directions are intended to broaden the potential relevance of the developed framework and, ultimately, to contribute to more robust and flexible anomaly detection capabilities in challenging real-time settings.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Nicholas Jeffrey, Qing Tan, and José R. Villar. (2024). A hybrid methodology for anomaly detection in Cyber–Physical Systems. *Neurocomputing* 568;1-7. https://doi.org/10.1016/j.neucom.2023.127068

[2] ABDULMALIK SHEHU YARO, FILIP MALY, PAVEL PRAZAK, AND KAREL MALÝ. (2024). Outlier detection performance of a modified Z-score method in time-series RSS observation with hybrid scale estimators. *IEEE*. 12;12785 - 12796. http://DOI:10.1109/ACCESS.2024.3356731

[3] Maha Shabbir, Sohail Chand, and Farhat Iqbal. (2024). Novel hybrid and weighted ensemble models to predict river discharge series with outliers. *Kuwait Journal of Science* 51(2);1-11. https://doi.org/10.1016/j.kjs.2024.100188

[4] Zhichao Hu, Xiangzhan Yu, Likun Liu, Yu Zhang, and Haining Yu. (2024). ASOD: an adaptive stream outlier detection method using online strategy. *Journal of Cloud Computing* 13(120);1-20. https://doi.org/10.1186/s13677-024-00682-0

[5] Franz Kevin Stehle, Wainer Vandelli, Giuseppe Avolio, Felix Zahn, and Holger Fröning. (2024). DeepHYDRA: A Hybrid Deep Learning and DBSCAN-Based Approach to Time-Series Anomaly Detection in Dynamically-Configured S. *ACM*, 272-285. https://doi.org/10.1145/3650200.3656637

[6] Mutasem K. Alsmadi, Malek Alzaqebah, Sana Jawarneh, Ibrahim ALmarashdeh, Mohammed Azmi Al-Betar, Maram Alwohaibi, Noha A. Al-Mulla, Eman AE Ahmed, and Ahmad AL Smadi. (2024). Hybrid topic modeling method based on dirichlet multinomial mixture and fuzzy match algorithm for short text clustering. *Journal of Big Data* 11(68);1-21. https://doi.org/10.1186/s40537-024-00930-9

[7] Maha Nssibi, Ghaith Manita, Amit Chhabra, Seyedali Mirjalili, andOuajdi Korbaa. (2024). Gene selection for high dimensional biological datasets using

hybrid island binary artificial bee colony with chaos game optimization. *Artif Intell Rev*.57(51);1-74. https://doi.org/10.1007/s10462-023-10675-1

[8] Gouranga Duari, and Rajeev Kumar. (2024). Attribute Subspace Partitioning with Neural Regression for Contextual Outlier Detection. *Procedia Computer Science* 235, pp.1892-1902. https://doi.org/10.1016/j.procs.2024.04.180

[9] ZHICHAO XIE, and XUAN HUANG. (2024). A Credit Card Fraud Detection Method Based on Mahalanobis Distance Hybrid Sampling and Random Forest Algorithm. *IEEE*, 1-15. http://DOI:10.1109/ACCESS.2024.3421316

[10] Dexun Jiang, Hao Zhu, Jie Liu, Xiaoxiao Feng, Fangjingxin Ma, and Jing Wang. (2024). Dynamic surface river pollution identification by a hybrid multivariate-based anomaly detection algorithm. *Journal of Cleaner Production*. 467;1-9. https://doi.org/10.1016/j.jclepro.2024.142923

[11] Gábor Princz, Masoud Shaloo, and Selim Erol. (2024). Anomaly Detection in Binary Time Series Data: An unsupervised Machine Learning Approach for Condition Monitoring. *Procedia Computer Science* 232;1065-1078. https://doi.org/10.1016/j.procs.2024.01.105

[12]Omar alghushairy, raed alsini, zakhriya alhassan, abdulrahman a. alshdadi, ameen banjar, ayman yafoz, and xiaogang ma. (2024). An Efficient Support Vector Machine Algorithm based Network Outlier Detection System. *IEEE*. 12;24428 - 24441. http://DOI:10.1109/ACCESS.2024.3364400

[13] Hugo M. Ferreira, David R. Carneiro, Miguel A. Guimar ˆ aes, and Filipe V. Oliveira. (2024). Supervised and unsupervised techniques in textile quality inspections. *Procedia Computer Science* 232;426-435. https://doi.org/10.1016/j.procs.2024.01.042

[14] Paul D. Rosero-Montalvo, Zsolt István, Pınar Tözün, and Wilmar Hernandez. (2024). Hybrid anomaly detection model on trusted IoT devices. *IEEE*. 10(12);10959-10969. http://DOI:10.1109/JIOT.2023.3243037

[15] Muhammad Ali, Peimin Zhu, Ma Huolin, Heping Pan, Khizar Abbas, Umar Ashraf, Jar Ullah, Ren Jiang, and Hao Zhang. (2024). A novel machine learning approach for detecting outliers, rebuilding well logs, and enhancing reservoir characterization. *Nat Resour Res* 32, 1047–1066. https://doi.org/10.1007/s11053-023-10184-6

[16] Henrique O. Marques, Lorne Swersky, Jörg Sander, Ricardo J. G. B. Campello, and Arthur Zimek. (2023). On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection. *Data Min Knowl Disc*. 37;1473–1517. https://doi.org/10.1007/s10618-023-00931-x

[17] Yajie cui, zhaoxiang liu, and shiguo lian. (2024). A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE*. 11;55297 - 55315. http://DOI:10.1109/ACCESS.2023.3282993

[18] Md Amirul Islam, Md Ashraf Uddin, Sunil Aryal, and Giovanni Stea. (2024). An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. *Journal of Information Security and Applications* 78;1-21. https://doi.org/10.1016/j.jisa.2023.103618

[19] K. Samunnisa, G. Sunil Vijaya Kumar, and K. Madhavi. (2023). Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods. *Measurement: Sensors* 25;1-12. https://doi.org/10.1016/j.measen.2022.100612

[20] Robert K. L. Kennedy, Zahra Salekshahrezaee, Flavio Villanustre, and Taghi M. Khoshgoftaar. (2023). Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. *J Big Data* 10(106);1-20. https://doi.org/10.1186/s40537-023-00750-3

[21] Ch. Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, Ashish Ghosh. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal* 6;1-8. https://doi.org/10.1016/j.dajour.2023.100164

[22] Rasha ramadan z. koko, inas a. yassine, manal abdel wahed, june k. madete, and muhammad a. rushdi. (2023). Dynamic construction of outlier detector ensembles with bisecting k-means clustering. *IEEE*. 11;24431-24447. http://DOI:10.1109/ACCESS.2023.3252004

[23]Ahsnaul Haque, Md Naseef-Ur-Rahman Chowdhury, Hamdy Soliman, Mohammad Sahinur Hossen, Tanjim Fatima, and Imtiaz Ahmed. (2023). Wireless sensor networks anomaly detection using machine learning: a survey. *Springer*, pp.1-21.

[24] Kyung sung lee , seong beom kim, and hee-woong kim. (2023). Enhanced Anomaly Detection in Manufacturing Processes through Hybrid Deep Learning Techniques. *IEEE*. 11;93368 - 93380. http://DOI:10.1109/ACCESS.2023.3308698

[25]Miloˇs Savi´c, Jasna Atanasijevi´c, Duˇsan Jakoveti´c, and Nataˇsa Kreji´c. (2021). Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method. *Expert Systems with Applications* 193;1-35. https://doi.org/10.1016/j.eswa.2021.116409

[26]David velásquez, enrique pérez, xabier oregui, arkaitz artetxe, jorge manteca, jordi escayola mansilla, mauricio toro, mikel maiza, and basilio sierra. (2022). A hybrid machine-learning ensemble for anomaly detection in real-time industry 4.0 systems. *IEEE*. 10,72024 - 72036. http://DOI:10.1109/ACCESS.2022.3188102

[27] Jian Zheng, Jingyi Li, Cong Liu, Jianfeng Wang, Jiang Li, and Hongling Liu. (2022). Anomaly detection for high-dimensional space using deep hypersphere fused with probability approach. *Complex Intell. Syst*. 8,4205–4220. https://doi.org/10.1007/s40747-022-00695-9

[28] Alona Sakhnenko, Corey O'Meara, Kumar J. B. Ghosh, Christian B. Mendl, Giorgio Cortiana, and Juan Bernab´e-Moreno. (2021). Hybrid classical-

quantum autoencoder for anomaly detection. *Springer*, pp.1-17.

[29] Andrey Kharitonov, Abdulrahman Nahhas, Matthias Pohl, and Klaus Turowski. (2022). Comparative analysis of machine learning models for anomaly detection in manufacturing. *Procedia Computer Science*. 200(0);1288-1297. https://doi.org/10.1016/j.procs.2022.01.330

[30] Lejla Begic Fazlic, Ahmed Halawa, Anke Schmeink, Robert Lipp, Lukas Martin, Arne Peine, Marlies Morgen, Thomas Vollmer, Stefan Winter, and Guido Dartmann. (2022). A Novel Hybrid Methodology for Anomaly Detection in Time Series. *Int J Comput Intell Syst*. 15(50);1-16. https://doi.org/10.1007/s44196-022-00100-w

[31] Bhanu Chander, and G. Kumaravelan. (2022). Outlier detection strategies for WSNs: A survey. *Journal of King Saud University - Computer and Information Sciences* 34(8);5684-5707. https://doi.org/10.1016/j.jksuci.2021.02.012

[32] LIWEN ZHOU, QINGKUI ZENG, AND BO LI. (2022). Hybrid anomaly detection via multihead dynamic graph attention networks for multivariate time series. *IEEE*. 10,40967 - 40978. http://DOI:10.1109/ACCESS.2022.3167640

[33] Thudumu, Srikanth; Branch, Philip; Jin, Jiong; Singh, Jugdutt (Jack). (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7(1),1-30. http://doi:10.1186/s40537-020-00320-x

[34] Wang, Biao; and Mao, Zhizhong . (2019). Detecting outliers in industrial systems using a hybrid ensemble scheme. *Neural Computing and Applications*, pp.1-17. http://doi:10.1007/s00521-019-04307-5

[35] Kurt, Mehmet Necip; Yilmaz, Yasin; and Wang, Xiaodong. (2020). Real-Time Nonparametric Anomaly Detection in High-Dimensional Settings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7);2463 - 2479. http://doi:10.1109/TPAMI.2020.2970410

[36] Xian-Fang Song; Yong Zhang; Dun-Wei Gong; and Xiao-Zhi Gao;. (2021). A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data. *IEEE Transactions on Cybernetics*, 52(9);9573 -9586. http://doi:10.1109/tcyb.2021.3061152

[37] Mohammed Qaraad; Souad Amjad; Ibrahim I. M. Manhrawy; Hanaa Fathi; Bayoumi Ali Hassan; and Passent El Kafrawy;. (2021). A Hybrid Feature Selection Optimization Model for High Dimension Data Classification. *IEEE Access*, 9;42884 - 42895. http://doi:10.1109/access.2021.3065341

[38] Yuan, Zhong; Chen, Hongmei; Li, Tianrui; Liu, Jia; and Wang, Shu . (2020). Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection. *Fuzzy Sets and Systems*, 421;1-28. http://doi:10.1016/j.fss.2020.10.017

[39] Chen, Gang; Du, Linlin; and An, Baoran . (2020). [IEEE 2020 Chinese Control And Decision Conference (CCDC) - Hefei, China (2020.8.22-2020.8.24)] *2020 Chinese Control And Decision Conference (CCDC) - Ordinal Outlier Algorithm for Anomaly Detection of High-Dimensional Data Sets.* Pp.5356–5361. http://doi:10.1109/CCDC49329.2020.9164610

[40] Yan Qiao; Kui Wu; and Peng Jin;. (2021). Efficient Anomaly Detection for High-Dimensional Sensing Data with One-Class Support Vector Machine. *IEEE Transactions on Knowledge and Data Engineering*, 35(1), pp. 404 - 417. http://doi:10.1109/tkde.2021.3077046

[41]C, A., K, S., N, N. S., & S, P. (2024). Secured Cyber-Internet Security in Intrusion Detection with Machine Learning Techniques. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.491

[42]K. Tamilselvan, , M. N. S., A. Saranya, D. Abdul Jaleel, Er. Tatiraju V. Rajani Kanth, & S.D. Govardhan. (2025). Optimizing data processing in big data systems using hybrid machine learning techniques. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.936

[43]Mekala, B., Neelamadhab Padhy, & Kiran Kumar Reddy Penubaka. (2025). Brain Tumor Segmentation and Detection Utilizing Deep Learning Convolutional Neural Networks: Enhanced Medical Image for Precise Tumor Localization and Classification. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1051

[44]S. Ranjana, & A. Meenakshi. (2025). Breast Cancer Detection using Convolutional Autoencoder with Hybrid Deep Learning Model. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1225

[45]Fowowe, O. O., & Agboluaje, R. (2025). Leveraging Predictive Analytics for Customer Churn: A Cross-Industry Approach in the US Market. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.20

[46]Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.19

[47]Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.18