A Learning-Based Ensemble Algorithm with Optimal Selection for Outlier Detection

Girish Reddy Ginni[1*], Srinivasa L . Chakravarthy[2]
[1]Department of Computer Science and Engineering, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh, India
girishloshankar@gmail.com, chakri.ls@gmail.com
[*]Corresponding author
Orcid: [1]https://orcid.org/0009-0005-5242-8839, [2]https://orcid.org/0000-0001-9141-4863

Keywords: ensemble outlier detection, learning based outlier detection, outlier method selection strategy, machine learning

Received: October 26, 2024

In this paper, we propose a Learning-based Ensemble Method with Optimal selection strategy (LbEM-OSS), which presents a new outlier detection algorithm that captures only outstanding ones of constituent models. Using KNN to define local regions and Pearson correlation to evaluate the detectors makes the ensemble robust. Our method can adapt and generalize better across different high-dimensional datasets by generating pseudo-ground truths with average and maximum aggregation strategies. On a wide range of benchmark datasets, LbEM-OSS outperformed both statistics-based and neural ensemble methods, which achieved state-of-the-art ROC-AUC as high as 97.78% in the best-case and 4-8% AUC improvements over existing methods on average. These results portray its potential for noise, different dimensionality, and heterogeneous data nature. Moreover, it is highly scalable and accurate, which makes it an essential application in practical fields like fraud detection, network security, and healthcare. This research highlights the need for dynamic selection approaches within ensemble methods, providing the groundwork for future developments in sound outlier detection.

Povzetek: Nova metoda ansambla, ki temelji na učenju, optimizira zaznavanje izstopov z dinamičnim izbiranjem najbolj zmogljivih modelov. Izboljša robustnost v visokodimenzionalnih naborih podatkov, s čimer doseže najsodobnejšo natančnost in razširljivost za aplikacije pri odkrivanju goljufij, varnosti omrežja in zdravstvenem varstvu.

1 Introduction

Outlier detection, a necessary part of data analysis, allows the identification of data points with significantly different values than the rest of the data set. Let's start with outliers. They can corrupt statistical studies and machine learning models; if they are not correctly handled, they can lead to incorrect results. Several statistical methods exist for outlier detection, including the Z-score, Tukey's fences, and isolation forests. For instance, anomaly detection is essential in the finance industry, fraud detection, and healthcare [1], [2]. Ensemble methods are necessary for improving the accuracy and robustness of outlier detection schemes, especially in high-dimensional data scenarios, as many features cause the problem of dimensionality curse and lead to higher complexity and noise than conventional outlier identification methods usually obtain robustness. To mitigate this problem, ensemble techniques combine several outlier detection algorithms to provide a final prediction that is more reliable and accurate. Ensemble methods reduce the impact of the shortcomings of individual models and offer a more holistic assessment of outliers in high-dimensional data by aggregating the results of several models [4], [5], [6]. Existing literature suggests that constituent outlier detection models play a crucial role in shaping an ensemble-based method's effectiveness for multiple reasons [10]. These models help detect and treat outliers in the dataset to avoid damaging the performance of the machine learning models. Ensemble models aggregate their predictions, which allows them to find outliers and minimize their effect, influencing the total accuracy. Moreover, the individual models may be over-fitted, and outliers alter the noise, but including the robust outlier-detection models in the ensemble strengthens them against noisy data. Additionally, by concentrating on the most significant data points, outlier detection encourages improved generalization, allowing the model to learn from representative instances, which enhances its forecasting capabilities on unseen data. Moreover, these models have provided additional information on the outliers' features, which helps improve the ensemble model's interpretability. As such, constituent outlier detection models are essential components of ensemble-based methodologies, leading to more accurate, robust, well-generalizing, and transparent insights into the data.

Ensemble methods are a new and powerful tool in the outlier detection arsenal, combining multiple models to harness each model's strengths to produce robust and accurate results. K-nearest neighbors (kNN) and Local Outlier Factor (LOF) are some of the key algorithms used in such methods. The kNN algorithm is effective for describing an isolated area using only k-nearest data points since it centers more on the local neighborhood structure of the data, which is especially helpful for anomaly detection. Similarly, LOF assesses the local density of a point compared to its neighbors and marks points with a substantially lower local density as outliers. They are complementary in that each method provides a solution for the diversity problem of the datasets with different characteristics, and integrating those methods into ensemble frameworks is the foundation of the proposed approach.

This paper presents the following contributions: We introduce a new algorithm named the Learning-based Ensemble Method with Optimal Selection Strategy (LbEM-OSS), focusing on the dynamic selection of top-performing constituent models to achieve a more robust outlier detection result. This would ensure that separate methods such as K-Nearest neighbors (KNN)-based models, Isolation Forests, and statistical outlier detection methods are rated based on their local and global relevance for being part of an ensemble. We use average and maximum aggregation strategies to generate pseudo-ground truths for this empirical evaluation. By calculating global and local ground truths, our algorithm reaches better accuracy, further improving adjustment to various high-dimensional datasets. The algorithm's performance is validated by several empirical studies conducted on benchmark datasets (re0, Sun09, Shuttle), and the algorithm produces the highest AUC score of 97.78%. Our proposed method can be used for fraud detection, healthcare, and network security applications and provides a reliable automatic outlier detection algorithm for complex data environments. The rest of the paper is organized as follows: The introduction summarizes the literature on various ensemble methods and information acquisition strategies. In Section 3, we propose an outlier detection algorithm. Section 4 examines the proposed method in experiments performed on some high-dimensional datasets. Section 5 presents a conclusion and suggests future research avenues.

2 Related work

In this section, research on existing ensemble learning approaches for outlier detection. Chakraborty et al. [1] proposed an innovative approach to outlier detection by integrating probabilistic neural networks and layered autoencoders, particularly addressing scenarios with multiple outliers and class imbalance. This method enhances detection accuracy by leveraging deep learning techniques for robust feature extraction. However, the study lacks dynamic selection mechanisms tailored for dataset-specific characteristics addressed in our proposed LbEM-OSS algorithm through KNN-based local regions and Pearson correlation evaluation. Reunanen et al. [2] suggested maximizing the selectivity and efficiency of outlier detection ensembles by using fewer instances. Our method adjusts parameters to yield a wide range of precise outcomes, which is advantageous for different algorithms. Boukerche et al. [3], significant research has addressed different difficulties over the last ten years by concentrating on effective outlier identification strategies. We classify new techniques, review their features, benefits, and drawbacks, and look at possible future developments. Zhong et al. [4] state that network traffic anomaly detection is essential for network security, but current approaches have problems with complexity, flexibility, and retraining. HELAD surpasses others by using deep learning algorithms—Abbasi et al. [5] required due to the increased data flow by flexible solution. ElStream outperforms traditional techniques in the detection of idea drifts using ensemble learning.

Fitriyani et al. [6] introduced an ensemble learning model for predicting diabetes and hypertension, integrating the system into a smartphone app for real-time diagnosis. While achieving high accuracy, the study focused on supervised ensemble methods and did not address unsupervised or semi-supervised scenarios common in outlier detection. Our approach builds on this by enhancing unsupervised ensemble techniques for high-dimensional datasets, thus broadening applicability. Schubert et al. [7] proposed a generalized outlier detection framework using flexible kernel density estimates, enabling the identification of anomalies in diverse data distributions without relying on rigid assumptions. Their approach demonstrates robustness in high-dimensional datasets, making it particularly relevant for ensemble-based outlier detection methods that enhance accuracy and adaptability. Zhang et al. [8] described utilizing stacking ensemble learning and multi-dimensional feature fusion in MFFSEM for intrusion detection. MFFSEM works better on a variety of datasets than current approaches. Li et al. [9] investigated Ps prediction using a Ps dataset and dimensionality reduction using four ensemble approaches. The results show the advantage of ensemble techniques. Zhu et al. [10] suggested a method for detecting intrusions on Internet of Things networks that combines ensemble learning with subspace clustering. It performs better than current techniques, with few false positives and excellent accuracy.

Ouyang et al. [11] improved machine learning analysis efficiency by introducing an EBOD approach for real-world datasets. Zhang et al. [12] presented DELR, a double-level ensemble approach to anomaly detection that aims to improve generalization capacity by tackling diversity and information loss. Suggested future enhancements include deep learning integration and real-time optimization, and DELR beats state-of-the-art algorithms on real-world datasets. Wang and Mao [13] addressed issues in process monitoring by introducing a dynamic ensemble outlier identification approach with one-class classifiers. Rigorous studies show its usefulness and future studies might focus on potential enhancements. Wang and Mao et al. [14] addressed ensemble difficulties by putting forth a dynamic outlier identification strategy that makes use of one-class classifiers. Experimental data demonstrate its efficacy over static ensembles and single models. The goal of more studies is to close the oracle's performance gap. Aljame et al. [15] used standard blood testing; early diagnosis of COVID-19 is essential. Combining classifiers improves predictions in an ensemble model called ERLX. The diversity of datasets and model validation continues to be challenged.

Zhang et al. [16] enhanced using a hybrid ensemble model that combines balanced sampling and outlier identification. In terms of prediction, it does better than benchmarks. Mienye et al. [17] improved across various domains through ensemble learning, integrating predictions from numerous models. Popular algorithms, including XGBoost and Random Forest, as well as bagging, boosting, and stacking techniques, are covered in this review. Yin et al. [18] employed 246 data sets and the stacking approach of ensemble learning. Outlier management and dimension reduction were incorporated into the preprocessing. An eight-model comparison revealed the advantage of ensemble models, mainly when dealing with skewed data. Tsai and Lin [19] evaluated 55 datasets to solve imbalanced class learning. OCC ensembles enhance performance. The influence of feature selection and multi-class unbalanced datasets will be investigated in future studies. Bull et al. [20] compared to supervised approaches, outlier ensembles perform comparably in damage identification and dimension reduction. Real-world engineering examples show how effective they are.

Zhang et al. [21] used for credit scoring have been altered by AI. In addition to improving feature interpretability and automatically optimizing parameters, a unique ensemble model handles outliers. Subudhi and Panigrahi [22] suggested a database security-focused intrusion detection system that combines OPTICS clustering and ensemble learning. Empirical findings demonstrate its advantages. Eddine et al. [23] used feature engineering and an RF classifier to create an intrusion detection model with excellent accuracy for IIoT security. Rovetta et al. [24] identified potentially dangerous occurrences in road audio streams. A novel ensemble approach combining one-class SVM for outlier identification and DNN for event classification shows promise. Cheng et al. [25] improved efficiency and accuracy with a two-layer ensemble technique that combines the Local Outlier Factor (LOF) for precise outlier identification with Isolation Forest (iForest) for rapid scanning and trimming.

Hus et al. [26] suggest that a stacked ensemble ANIDS using AE, SVM, and RF models be used for network intrusion detection. Tested on actual campus logs, NSL-KDD, and UNSW-NB15 datasets, it performs better than conventional models, decreasing incorrect predictions. Wei et al. [27] defend against adversarial assaults and out-of-distribution inputs. XEnsemble, a technique for DNN models, combines input and output verification. Biswas and Samanta [28], with Decision Tree, Naive Bayes, and kNN as essential learners, use ERF to address finding anomalies in wireless sensor networks. The AReM dataset evaluation reveals that ERF performs better than individual learners. In the future, multi-class categorization could be used. Jiang et al. [29], outlier identification in the Internet of Things is challenging because of resource limitations and wireless transmission. With an emphasis on their performance and unresolved research concerns, this review contrasts machine learning-based methods for outlier detection. Tsogbaatar et al. [30] used SDN to anticipate device state, manage flows, and identify abnormalities; the deep ensemble learning framework DeL-IoT tackles IoT risks.

Table 1: Comparative analysis of existing ensemble outlier detection methods with their performance metrics and limitations

Ref	**Methodology**	**Dataset**	**Performance (AUC/Precision)**	**Limitation**
[1]	Probabilistic neural networks with layered autoencoders	Custom dataset	AUC: 85.3%	Limited scalability and lacks adaptive detector selection.
[2]	Outlier detection ensemble	UNSW-NB15	AUC: 82.1%	Suboptimal feature selection and fixed detector configurations.
[3]	Deep learning ensembles	IoT datasets	Precision: 89.5%	Limited ability to address class imbalance and complex outliers.
[10]	Subspace clustering with ensembles	UNSW-NB15	AUC: 88.4%	Hyperparameter tuning challenges and narrow dataset focus.

Proposed	LbEM-OSS	Multiple high-dimensional datasets	**AUC: 97.78%**	Outperforms existing methods in adaptability and robustness.

Table 2: Summary of literature findings

Ref.	Approach	Technique	Algorithm	Dataset	Limitation
[2]	Deep Learning and Machine Learning	Outlier detection ensemble	outlier detection algorithms	Custom dataset	Further work will need to experiment with other optimization methodologies.
[10]	Bottom-up and Threshold-based approach	ML and Anomaly-based techniques	Clustering algorithms, namely CLIQUE, PROCLUS, and SUBCLU	UNSW-NB15 dataset	Upcoming projects will improve feature selection skills, refine hyperparameter tuning, evaluate the approach on different datasets and real-world situations, and guarantee the method's morally and practically sound implementation.
[16]	Machine learning and ensemble learning	Clustering and hyper-parameter optimization techniques	classic outlier detection algorithms	the UC Irvine (UCI)	Future research should consider and appropriately avoid any potentially negative behaviors and discriminatory practices of artificial intelligence systems toward humans.
[17]	Machine learning and Ensemble Learning	Ensemble and blending techniques	state-of-the-art algorithms and ensemble algorithms	European cardholders' dataset and Brazilian credit dataset,	It is thus advised that ensemble clustering be the subject of future study.
[23]	Ensemble Learning, DL, and ML	ML and encryption techniques	ML algorithms	Bot-IoT and NF-UNSW-NB15-v2 datasets.	Our future work will utilize other datasets, such as the TON-IoT dataset comprising IoT and IIoT data, to gain a worldwide perspective and develop and evaluate an efficient IDS for enhancing network security.
[24]	Ensemble Outlier Detection Approach	cutting edge methodologies	SVM and clustering algorithms	Custom dataset	The suggested approach will be improved to recognize events even when background noise heavily distorts their signals.
[28]	Density-based approach	Machine Learning techniques	ERF algorithm	Intel Berkeley Research lab (IRLB) dataset	Determining the many stages of nature may be our future direction when utilizing multi-class classifiers.
[32]	Outlier detection approach	RSS based techniques	A clustering-based outlier detection algorithm	UCI dataset	Future research might examine the suitability of artificial intelligence (AI) methods for outlier identification in localization and wireless sensor networks (WSNs).
[35]	ANNODE approach	Machine Learning techniques	SVM algorithms (H-OCSVM and QS-OCSVM)	Intel Berkeley Research Lab Mica2dot dataset	Future work will define specific methods (such as offset and drift) for continuous failure detection and expand the assessment to include more sensors.
[39]	Step-by-step pharmaceutical treatment approach	Machine Learning techniques	Ensemble learning and supervised learning algorithms	Custom dataset	The following are a few goals that might be explored in further research: 1. Increasing the performance of asthma control level detection by including additional elements, such as genetic factors and biomarkers, that impact asthma control levels. 2. Relying on time series analysis to implement asthma control

					level detection models instead of attribute-based data. 3. Using new technologies to develop self-care systems using models for detecting asthma control levels.

Table 3: An overview of the data sets used in the literature

Dataset (s)	**References**
UNSW-NB15 dataset	[10]
UCI	[16],[32]
European cardholders dataset and Brazilian credit dataset	[17]
Bot-IoT and NF-UNSW-NB15-v2 datasets.	[23]
Intel Berkeley Research lab (IRLB) dataset	[28],[35]
Custom dataset	[2],[24], [39]

Belhadi et al. [31] offered a model that outperforms current techniques in detecting anomalous human behavior by utilizing data mining and deep learning technologies. Bhatti et al. [32] presented "iF_Ensemble," a Wi-Fi indoor localization technique that combines ensemble, unsupervised, and supervised approaches. By detecting outliers, accuracy is increased by 2%. Wang et al. [33] identified the shortcomings of the existing iNNE architecture for wireless sensor network outlier identification, including flexibility and resource usage. Khare et al. [34] investigated the use of ensemble ML for anomaly detection in IoT contexts and compared it to conventional techniques. Jesus et al. [35] suggested machine learning, namely ANNODE, to identify reliable outliers in environmental sensor networks. It has been verified using actual datasets and has outperformed competing solutions.

Liu et al. [36] addressed sparsity in high-dimensional data by introducing SO-GAAL for outlier detection. This strategy is expanded by MO-GAAL, which outperforms rivals on a range of datasets. Xu and Chen [37] suggested a unique approach to anomaly detection for GSHP systems that combines statistical modeling and deep learning. Anomalies found are classified and verified, demonstrating the efficacy of the approach. Future studies will improve the evaluation of anomaly severity. Kapucu et al. [38] suggested a way for photovoltaic systems to diagnose faults using ensemble learning that increases generalization and classification accuracy. Khasha et al. [39] determined asthma control levels, a revolutionary ensemble learning technique that integrates machine learning algorithms with the experience of clinicians. Chai et al. [40] use human input, human-in-the-loop outlier detection, or HOD, to detect outliers precisely. To reduce human labor, HOD uses a bipartite graph-based technique with clustering to provide context inliers. Experimental data confirm the advantage of HOD. The literature showed a need to develop the best selection strategy for finding constituent outlier detection models to be part of an ensemble approach. Breunig et al. [41] introduced the LOF (Local Outlier Factor) method, which identifies outliers based on the local density deviation of a data point compared to its neighbors. This approach effectively handles varying densities within datasets, making it a foundational technique for local region-based outlier detection and ensemble methods leveraging neighborhood information. As seen in Table 1, existing ensemble methods often struggle with scalability, adaptive detector selection, and achieving high accuracy across diverse datasets. These limitations underline the necessity for our novel selection strategy, which integrates KNN-based local region definition and Pearson correlation for dynamic detector evaluation, achieving significantly higher performance. Table 2 summarizes the literature findings, while Table 3 presents the datasets used in the literature. The literature review observed that performance needs to be improved in selecting appropriate detection methods for ensemble models to enhance outlier detection effectiveness. Luo et al. (2021) developed a convolutional neural network (CNN) to autonomously identify acute ischemic stroke in brain magnetic resonance imaging (MRI) data. [42] Test results revealed that the designed model significantly outperformed the pre-improvement model in the social network data recommendation task. Choudhary et al. [43]

3 Proposed system

Unsupervised learning is key in identifying outliers and critical in multiple domains, such as fraud detection, network security, or quality assurance. It does not require labeled data and is usually used for clustering, density estimation, etc. It helps prevent fraud by identifying outliers, network security, and high-quality goods & services. Unsupervised outlier detection not only helps maintain the health and reliability of data-driven systems, but it also helps identify strange or dubious data points far from the mean. Outlier Detection – Clustering-based unsupervised learning is essential in outlier detection, as it identifies any data point that does not fit the expected pattern or any cluster of the data set. Using clustering to identify groups of similar data points makes it possible to identify outliers (i.e., single data points that do not belong to any cluster or single data points spread out amongst identified clusters). This technique helps find anomalies in massive data sets that are impossible to check manually. When we use hierarchical clustering to detect outliers, the result may consist of clustering algorithms like K-means, DBSCAN, and data points from outside the clusters, which are probably outliers. These exceptions may be mistakes in data assembly, fraud, or once-in-a-lifetime occurrences that may fascinate analysts. This clustering-based unsupervised learning approach can assist organizations in boosting the quality of their data, tighten up fraud detection, and allow organizations to unlock insights from unusual data points.

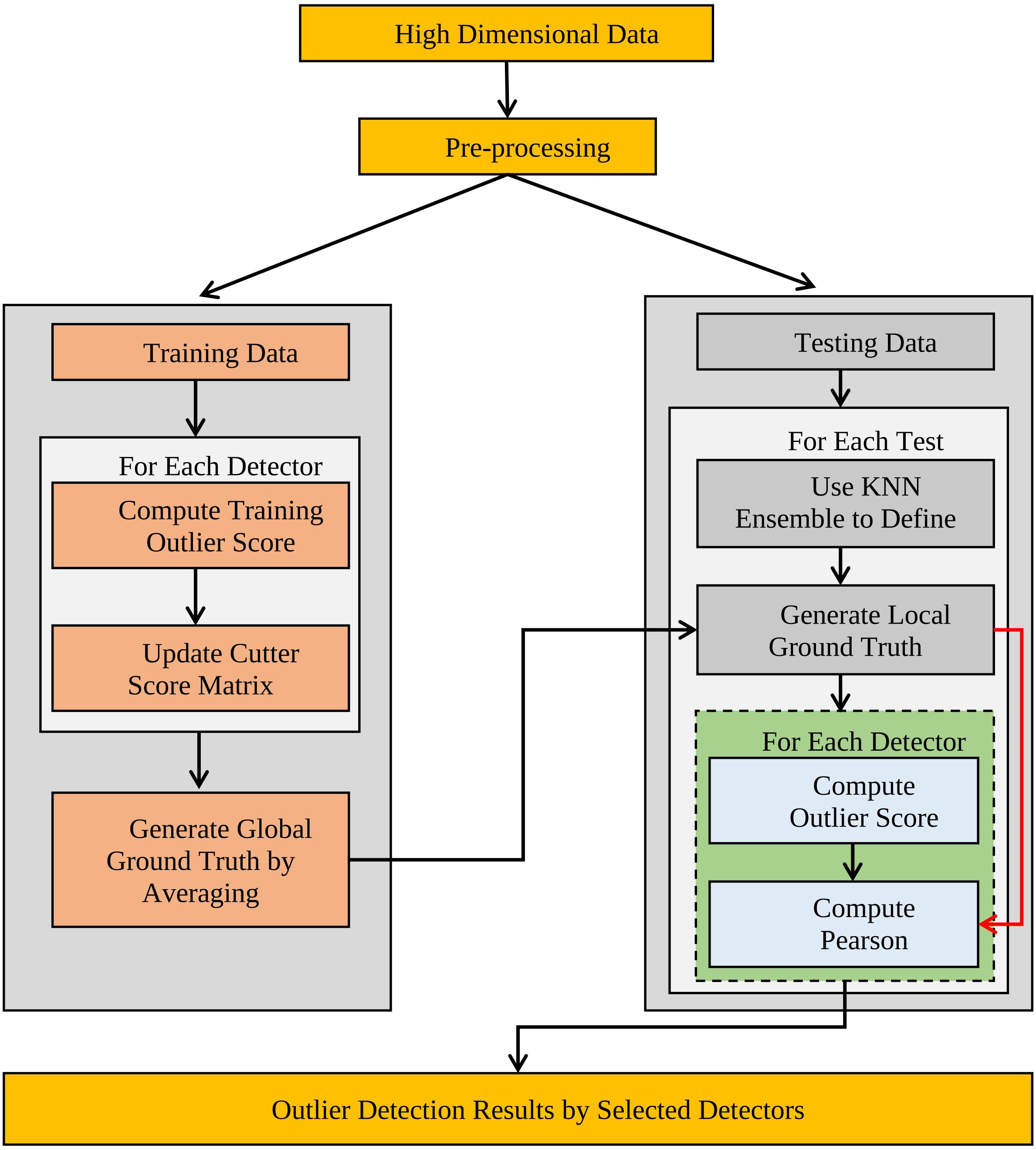


Figure 1: Ensemble learning-based framework

Figure 1 is the proposed ensemble learning-based framework to improve outlier detection performance. The process begins with high-dimensional data. Pre-processing is applied to the high-dimensional data. Each detector computes the training outlier score, updates the outlier score matrix, and creates a global ground truth by averaging using the training data. The testing data is utilized for every test instance throughout the testing process. Use the KNN ensemble to characterize the immediate area. Provide a ground truth for the area. For each detection, the framework computes the outlier score and computes the Pearson correlation. The outlier detection results are obtained from the selected detectors. In high-dimensional data, not all features are relevant for outlier detection. Preprocessing can include techniques to reduce dimensionality by eliminating redundant or irrelevant features. This increases computational efficiency and improves the accuracy of outlier detectors. Data may have different ranges and units. Normalizing and scaling the data ensures that all features contribute equally to calculating distances and similarities, which are critical in outlier detection. Incomplete data can introduce errors in the analysis. During preprocessing, methods for filling in the blanks or deleting partial records may be used to handle missing data, thus improving the dataset's quality. Data may contain noise that can confuse outlier detectors. Preprocessing can include filtering techniques to remove noise, ensuring that detectors focus on actual abnormalities in the data. In some cases, transforming the data to a different space can make underlying structures more evident and more accessible to detect as outliers. Preprocessing improves the quality and relevance of the data fed into the outlier detection framework, resulting in more accurate and efficient detection. Initially, our framework combines a set of diverse detectors. It first determines the vicinity of every test occurrence before selecting the best capable local detector (or detectors). The test instance's outlier score is produced using the chosen detector or detectors.

The framework relies on two key components: local pseudo-ground truth generation and dynamic outlier ensemble selection. Local pseudo-ground truth refers to the benchmark score computed for each test instance by aggregating the outlier scores of its k-nearest neighbors. This localized reference ensures context-aware evaluation of detector accuracy. Dynamic outlier ensemble selection is the process of adaptively selecting detectors with high correlation to the pseudo-ground truth, ensuring the ensemble is optimized for dataset-specific characteristics. These components enhance the framework's robustness and adaptability across diverse datasets.

3.1 Base detector generation

To encourage learning unique features in the data, a productive ensemble should be built with diverse base estimators [24], [32]. One way to introduce diversity among a collection of homogenous base detectors is to subsample the training set and area of features or adjust the model hyperparameters [6], [32]. By building a pool of models with the same fundamental technique utilizing different hyperparameters, we show how effective the framework is in this work. However, heterogeneous base detectors may also be employed with the proposed algorithm as a generic framework.

With n points and d features, the representation of the training data is $X_{train} \in R^{n\times d}$, while the representation of the test set is $X_{test} \in R^{m\times d}$ with m points. Initially, a collection of base detectors $C = \{C_1, \ldots, C_R\}$ is created by the method and populated with a variety of hyperparameters, such as a collection of LOF detectors with different MinPts [5]. The same dataset is used for inference once all base detectors have been trained on X_{train}. After combining the data, an outlier score matrix $O(X_{train})$, is created, which is represented by the score vector from the r^{th} base detector, Cr (•), in Eq. (1). Z-normalization is used to normalize each detector score $C_r(X_{train})$ by earlier research [2,32].

$$O(X_{train}) = [C_1(X_{train}), \ldots, C_r(X_{train})] \in R^{n\times d} \quad (1)$$

3.2 Pseudo ground truth generation

Two methods are used to create a false ground truth (denoted target) with $O(X_{train})$LSCP evaluates detector competency in the lack of ground truth labels. Average base detector scores (i) and maximum scores for all detectors (ii) are shown. This is further discussed in Eq. (2), which represents the entire aggregate (average or maximum) of all base detectors.

$$target = \emptyset(O(X_{train})) \in R^{n\times 1} \quad (2)$$

Note that the proposed system's fictitious ground truth was developed solely for detector selection and is based on training data.

3.3 Local region definition

The set of a test instance $X_{test}^{(j)}$'s k closest training objects is known as its local region, or ψ_j. Technically, this is indicated as:

$\psi_j = \{x_i \mid x_i \in X_{train,x_i} \in kNN_{ens}^{(j)}\}$ where the collection of a test instance's closest neighbors, as determined by an ensemble criterion, is described by kNN_{ens}. This kNN variant—which is comparable to Feature Bagging [1], suggested making use of kNNs for superior accuracy over clustering methods in DCS [9] and allaying worries about the curse of dimensionality on kNN [4]. The steps involved are as follows: Additional training objects are added if they occur more than t/2 times to $kNN_{ens}^{(j)}$ to define the local region. (i) t groups of $[\frac{d}{2}, d]$ to create new feature areas, features are chosen at random. (ii) The k nearest training objects to $X_{test}^{(j)}$ Use the Euclidean distance to identify each group. The region's size is not fixed since it depends on how many training items fulfill the selection requirements.

The number of closest neighbors to take into account throughout this procedure is determined by the local area factor k; excessive values are avoided. Larger values of k may focus too much on global connections and incur higher computational expenses; on the other hand, lesser values of k emphasize local links more, which may cause instability. While cross-validation [16] may be used to find an optimum k when ground truth is provided experimentally, there is no comparable simple method for unsupervised settings. Due to these factors, we advise using k = 0.1n, 10% of the training samples, with a restricted range of [30,100], which produced positive practical results.

3.4 Model selection and combination

To extract the local pseudo-ground truth $target^{\psi_j}$ for every test instance, retrieve values from the target that correspond to the local area ψ_j:

$$target^{\psi_j} = \{target_{x_i} | x_i \in \psi_j\} \in R^{|\psi_j| \times 1} \quad (4)$$

where the cardinality of ψ_j is indicated by $|\psi_j|$. In the same way, the pre-calculated training score matrix $O(X_{train})$ may be used to extract the local training outlier scores $O(\psi_j)$ as follows:

$$O(\psi_j) = [C_1(\psi_j), \ldots, C_r(\psi_j)] \in R^{|\psi_j| \times R} \quad (5)$$

In light of this, it is possible to effectively get the local outlier scores and targets from precalculated values, even if the local region must be recomputed for every test instance.

While the proposed system evaluates the similarity between base detector scores and the pseudo goal, DCS assesses the accuracy of base classifiers as the proportion of adequately classed points [16] for determining base estimator skill in a limited region. The reason behind this divergence is the lack of well-defined and consistent techniques for accessing binary labels in unsupervised outlier mining. Although converting pseudo-outlier scores to binary labels is feasible, choosing the suitable conversion threshold is challenging. Furthermore, using similarity measures rather than absolute accuracy for competency evaluation is more stable because outlier identification jobs often include imbalanced datasets. Consequently, LSCP calculates the local competence of every base detector by utilizing the local pseudo-ground truth's Pearson correlation. $target^{\psi_j}$and the local detector score $C_r(X_{train}^{\psi_j})$., which is helpful in outlier ensemble model combinations [25]. For $X_{test}^{(j)}$, the detector C_r^* with the highest similarity is deemed to be the most capable local detector; hence, its outlier score $C_r^*(X_{test}^{(j)})$ may be regarded as the test sample's final score.

3.5 Dynamic outlier ensemble selection

In unsupervised learning, choosing just one detector might be dangerous, even if it is the one that most closely resembles the pseudo-ground truth. One way to lower this risk is to select a group of detectors for a second-phase combination. This concept may be understood as a modification of supervised DES [16] for outlier identification; thus, we provide ensemble versions of the system that utilize the Average of the highest and lowest values of average ensembling techniques. More specifically, MOA selects a set of competent detectors near a test instance and, when ∅_average generates the pseudo ground truth, uses the maximum of their predictions as the outlier score. On the other hand, AOM determines the average of the selected subset when the pseudo target is created using ∅_max. Setting the group size of selected detectors to one is one instance when the ensembles provide the original algorithms. While a group size of R results in a genuinely global algorithm, more prominent group sizes can be considered more international in their detector selection. In light of this, we recommend using a variance-adjusted group size selection method. Specifically, a histogram of detector Pearson correlation scores (to the fictitious ground truth) is constructed using b equal intervals. The detectors from the most frequent interval are retained for the second-phase combination. Fewer detectors are chosen when b is significant, which flexibly regulates the group size strength in proposed ensembles.

The complexity of training each base detector and generating the pseudo-ground truth depends on the underlying model and the number of training samples. However, since this study suggests a combination structure, we concentrate on the overhead added at the combination step in our discussion. The additional time required to define each test instance's local area is $O(nd + nlog(n))$: $O(nd)$ for the distance computation and O(nlog(n)) for summing and sorting, with the proper implementation of the models, for example, using a k-d tree [16]. Here, n and d represent each test case and its dimensionality. Although defining the local region necessitates several rounds, the complexity analysis does not account for the fixed number of iterations. An extra O(s) is required to combine the s base detectors in MOA and AOM, resulting in an $O(nd + nlog(n) + s)$overall time complexity.

Aggarwal and Sathe recently employed the biased-variance trade-off, a popular approach for assessing erroneous generalization in classification problems, to lay the theoretical groundwork for outlier ensembles [2]. Where there is typically a trade-off between these two channels, squared bias or variance can be decreased to decrease the reducible generalization error in outlier ensembles. A high-bias detector may not perform well with complicated data, but it is less susceptible to data fluctuation than a high-variance detector in terms of instability. Controlling variance and bias is the aim of outlier ensembles to lower the total generalization error. This new approach has been used to assess several recently presented algorithms to improve interpretability [23, 24, 30].

It has been demonstrated that variance reduction occurs when diverse base detectors are combined, for example, by averaging them [2,23,24]. However, some of the base detectors in the mixture may be false, which would raise bias. This explains the poor performance of generic global averaging. The proposed system mixes variance and bias reduction in Aggarwal's bias-variance framework. Starting different base detectors with different hyperparameters to generate pseudo-ground truth generates diversity and subtly promotes variance reduction. The method also prioritizes local competency-based detector selection to help find base detectors with conditionally low model bias. The framework is also expected to be more stable than global maximization (GG_M) as the variance is reduced by using the output of the most competent detector instead of the global maximum of all base detectors. In their second phase, they significantly minimize generalization errors by reducing variance and bias.

3.6 Research design

Linking existing ensemble-based outlier detection methods, the research intends to fill in gaps by introducing an ensemble mechanism dubbed the Learning-based ensemble Method with Optimal Selection Strategy (LbEM-OSS). Provide a dynamic framework for ensemble design that enables the selection of high-performing outlier detection models based on local relevance; Use K-Nearest Neighbors (KNN) for defining local regions to ensure context-sensitive outlier detection; and establish a robust, accurate ensemble-based approach by leveraging Pearson correlation for detector evaluation and selection. Additionally, the study aims to assess the anticipated framework in high-dimensional datasets and compare its findings with the latest methods. This will be driven by at least a few key hypotheses associated with the study. The proposed LbEM-OSS algorithm should achieve improved accuracy and robustness compared to the ensemble outlier detection approaches. Our second assumption is that using KNN for local region identification and Pearson correlation for the detector will be more effective and will thus lead to higher performance values (mean average precision and AUC).

A critical parameter of the method is the size of the local region of each test instance, which is defined by the local area factor (k), significantly affecting the algorithm's performance. This parameter specifies the number of nearest neighbors" to be taken from where the local region is formed, thus substantially impacting the granularity of local ground truth, computational time, and, consequently, detection accuracy. More significant k values highlight global connections, which can wash out local traits necessary for outlier detection in a better and more reliable way, and k values less emphasize local linkages but are potentially unstable due to limited local characteristic representation. An empirical upper bound of k = 0.1 n k = 0.1n (10% of training samples) is found, restricted between 30 and 100, that achieves a good compromise between variability and computational burden. Choosing the best k is especially difficult in unsupervised setups as no label data is available. The suggested method advises testing a subset of the pseudo-ground truth data using cross-validation to find an appropriate k. The primary goals of this investigation, together with parameter considerations, align with the overall purpose of improving outlier detection methods. The variability of kk also makes the approach more effective in a broader array of datasets and scenarios. In future work, We will refine this approach by automatically optimizing k using some appropriate metaheuristic techniques.

3.7 Proposed algorithm

This article, A Learning-Based Ensemble Method with Optimal Selection Strategy Abstract Outlier detection strategies have been well-studied in low- or high-dimensional datasets. It starts by dividing the dataset into test and training sets and then calculating each detector's outlier scores using the training data. A global ground truth is generated, and a local region is defined for each instance in the test set to create a local ground truth. It chooses which detectors are most relevant to a particular ensemble by measuring the correlation between the score of each detector and the local ground truth. Last, it outputs a final outlier score for every instance based on the selected detectors so outlier detection results can be derived.

Algorithm: Learning-based Ensemble Method, with Optimal Selection Strategy (LbEM-OSS)

Input: High dimensional dataset D, candidate outlier detectors C, number of neighbors n, threshold th

Output: Outlier detection results R, performance statistics P

```
1.  Begin
2.  Initialize selected outlier detectors vector S
3.  (T1, T2)←DataSplit(D) //training and test data
4.  For each outlier detector candidate c in C
5.    score←getOutlierScore(c, T1) //computes outlier score
6.    matrix←updateOScoreMatrix(c, score)
7.  End For
8.  ggt←computeGGTruth(M)  //generation of global ground truth
9.  For each instance t in T2
10.   lregion←computeLocalRegion(n, t, T2) //compute local region using KNN
11.   lgt←computeLGTruth(ggt, lregion) //generation of local ground truth
12.   For each outlier detector candidate c in C
13.     score←getOutlierScore(c, lregion, T1) //computes outlier score
14.     pc←computePearson(score, lgt) //compute Pearson correlation
15.     IF pc>=th Then
16.       add c to S  //S has constituent outlier detection methods of ensemble
17.     End If
18.     foscore←computeFOScore(S, t) //compute final outlier score
19.     Add t and foscore to R
20.   End For
21. End For
22. Print R
23. End
```

Algorithm 1: Learning-based Ensemble Method, with Optimal Selection Strategy (LbEM-OSS)

Algorithm 1: Outlier Detection Algorithm uses multiple outlier detection algorithms to provide outlier detection in high dimensional datasets and improve efficiency. This fundamental idea of fusing complementary candidate detectors can be applied to an optimal selection scheme to improve the robustness and accuracy of detection pipelines. The algorithm initializes a vector(S)to hold selected outlier detectors for the ensemble. Then, Data is split (D) for train (T1) and test (T2). We can create two datasets based on our dataset, i.e., The training and testing datasets. This is a significant split as it will make the algorithm train the ensemble on 1 dataset. At the same time, the performance will be measured by taking a different testing dataset, which will not provide some bias as it will overfit the training data.

The algorithm will run each candidate outlier detector (c) in a set (C) and calculate the outlier score using the training data (T1). It provides a score indicating how likely each instance in the dataset will be an outlier based on the features that the detector (c) learned. The results will be the outlier score matrix, a base for further calculations. A consolidated score matrix from the individual candidate detectors produces a global ground truth based on this new score matrix. It is a reference against which local ground truths will be compared to gain a complete overview of the outlier structure over the entire dataset. The algorithm individually processes each instance (t) from the testing set(T2). The local region for every instance is computed using the k-nearest neighbors (KNN) method to perform the evaluation locally concerning the instance. Local ground truth is defined over this region using the global ground truth computed previously, enabling a focused review of outlier characteristics that may differ from the global perspective.

All candidate detectors (c) are validated based on the local ground truth in the next step. An outlier score is calculated for the instances in the local area, and the Pearson correlation coefficient between the calculated score and the local ground truth is measured. This correlation measures how much a detector agrees with the correct outlier status of the instances considered. If the correlation equals or exceeds L=ht, the detector (c) is part of the ensemble set (S). After we have chosen the best detectors, we calculate the final outlier score for each instance (t) based on all selected detectors in (S). This statistical score value thus represents the final metric for each example of whether they are to be considered abnormal or not. The output vector (R) contains the results (t, the final outlier score for t). The LbEM-OSS algorithm provides a principled framework for outlier detection in high-dimensional datasets. It aims at an optimal selection strategy of a unity set of several detectors such that the strengths of each method are maximized and the weaknesses of the methods are minimized. So, combining global ground truth with local ground truth enhances the correctness and reliability of the results from outlier detection [30], making this method more robust and applicable for data analyses across different fields compared to some of the methods above.

The local region of each test instance consists of k neighbors that arrive between the granularity and computational overhead of the k values. Although k=0.1n (10% of the dataset size) is the best value over several datasets empirically, for more minor data, k is set in the range of 30–100 to guarantee that sufficient neighbors are considered without substantial noise. This choice balances local sparsity and keeps the algorithm sensitive to the local context.

We used Pearson correlation as the evaluation metric for the detector competence since it can quantify the linear relationship between the detector score and the pseudo-ground truth. The scoring function also works well with the aggregation methods in the algorithm, e.g., averaging and maxing the individual detector scores, such that valuable detectors that highly correlate to these reference numbers are retained. Although alternatives such as Spearman correlation or mutual information might better describe non-linear or rank-based patterns, Pearson correlation was found to work better in this space empirically.

The generation of pseudo ground truth is a key component in the proposed Learning-based Ensemble Method with Optimal Selection Strategy (LbEM-OSS); we use two aggregation strategies for the pseudo ground truth: average and max scores overall base detectors. And this is why these selections are made — their benefits fit nicely together. While the average score might smooth out variations across detectors and provide a stable representation of potential outliers, the maximum score captures extreme values that can signify extreme outlier behavior. These strategies guarantee that the pseudo ground truth encompasses global trends and corner cases, enabling super effectiveness in the case of unsupervised learning.

Local area factor kk from equation (9) (throughout the paper, this factor is set to 0.1n0). The value of p, (1n(10% of training samples)), was estimated using brute force. It is a balance between emphasizing local relations and computational efficiency. It could describe smaller values of k, which is a more localized perspective. Still, as described in the Theory section, they may be unstable — too few points in a neighborhood may not capture the general essence. In comparison, global trends heavily influence larger values; thus, they may not help or even blur the available information on detecting the outliers successfully. Various k values (30 to 100) were tried for multiple datasets, and the best was n0. In all cases, 1n seemed to produce the best results. Future framework releases will investigate more sophisticated parameter optimization methods for this selection (for example, cross-validation over pseudo ground truth or meta-heuristic algorithms).

The computation overhead, in this framework, local regions must be recomputed at each test instance using K-Nearest Neighbors (KNN), which is expensive for high-dimensional datasets. The complexity of this process is O(nd) O(nd) when calculating the distance between each test instance and all training samples and O(nlog n) O(n log n) when sorting and summing, assuming implementation of efficient algorithms like k-d tree. The overall complexity increases by generating local ground truths and Pearson correlations for model detector selection. Although these steps allow one to detect outliers more accurately and context-relatively, they come with a trade-off regarding their run-time. In future work, we will thus try to reduce this computational overhead by running KNN through efficient algorithm approximations like locality-sensitive hashing (LSH) for increased scalability.

4 Experimental results

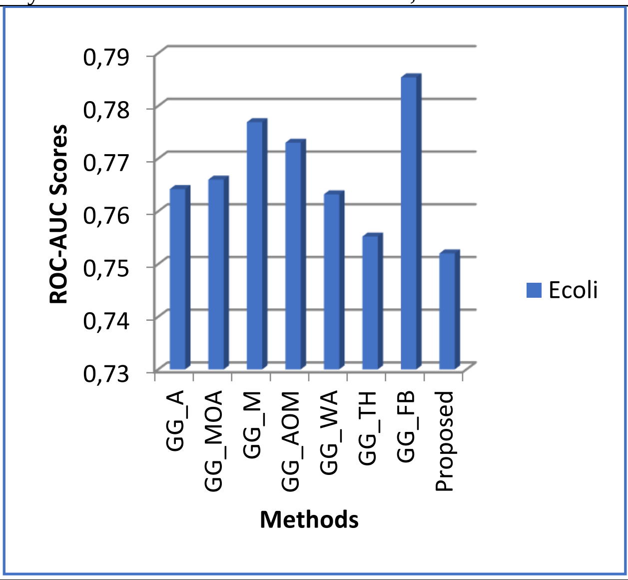
This section reports the results of a practical experiment implemented on different high-dimensional data sets. Furthermore, the results of the proposed approach are in contrast to numerous state-of-the-art outlier detection approaches. A comparison of ROC-AUC and mean average precision for all the outlier detection methods assessing the performance of outlier detection methods on identifying outliers

Tr23, Wap, Glass, Shuttle, Kddcup, Ecoli, Yeast, Caltech, Sun09, Fbis, K1b, re0, re1, Tr11, which are used for the evaluation datasets. These datasets cover a wide area of domains and difficulties: biological data (Ecoli and Yeast), computer vision (Caltech and Sun09), textual data (Tr23, Wap, Fbis, K1b, re0, re1, and Tr11), and network security (Shuttle and Kddcup). These were chosen to assess the generalization and stability of the LbEM-OSS algorithm as the data's nature is high dimension, class imbalanced, and noise.

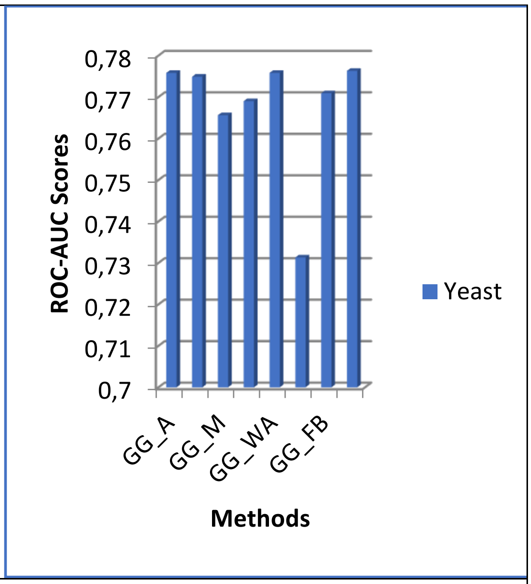
We empirically evaluate the proposed Learning-based Ensemble Method with Optimal Selection Strategy (LbEM-OSS) optimization framework over the high-dimensional datasets Ecoli, Yeast, Caltech, Sun09, etc. These datasets were selected according to their approximate representation for high-dimensional outlier detection tasks. In widely varied domains such as biology, computer vision, and network security, these datasets naturally contain high-dimensional data with complex structures and outliers. Ecoli and Yeast are standard bioinformatics datasets for genetic and protein data where anomalies present as differences in patterns exposed from everyday observations. At the same time, Caltech and Sun09 represent visual data datasets reflecting data analysis challenges where high-dimensional features can inhibit the identification of anomalies. Such variation in dataset characteristics guarantees the proposed method is robust and generalizable to different application scenarios.

We assess LbEM-OSS performance with well-known metrics, such as ROC-AUC and mean average precision (mAP). These measures are relevant for evaluating the performance of outlier detection approaches, showing the true positive rates and the precision-recall trade-off for different thresholds. Finally, we apply statistical tests, including paired t-tests and Wilcoxon signed-rank tests, to verify statistically that the performance differences between LbEM-OSS and baseline methods are statistically significant. They test in a statistical way if the gains we see are real and not just random noise in the data.

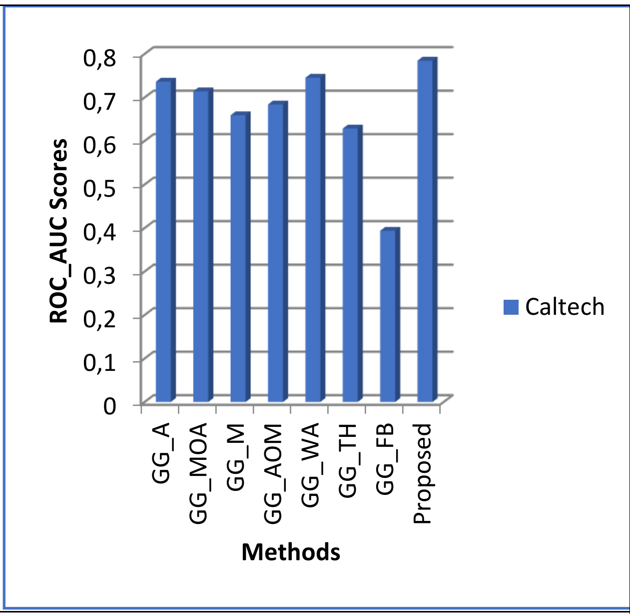
The statistical tests provide background to the quantitative results and add rigor to the comparative analysis. The proposed approach is reliable, e.g., the statistical significance of LbEM-OSS producing a much higher AUC score (e.g., 97.78%) compared to existing methods (e.g., subspace clustering ensemble giving 88.4%). The experimental design used in this paper provides a comprehensive assessment of LbEM-OSS's performance by combining different datasets, solid evaluation metrics, and statistical validation.



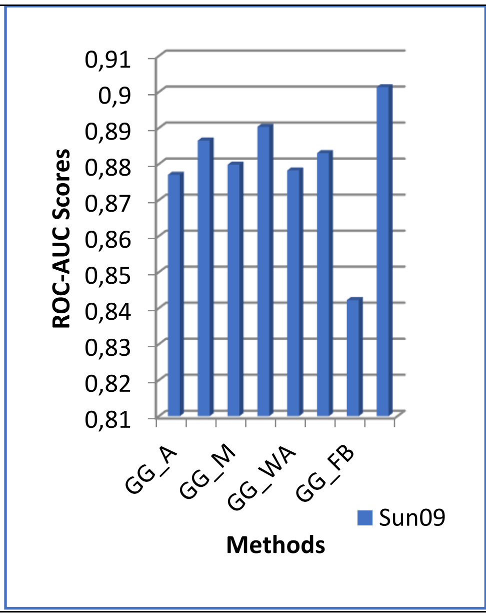
(a) Ecoli dataset



(b) Yeast dataset



(c) Caltech dataset

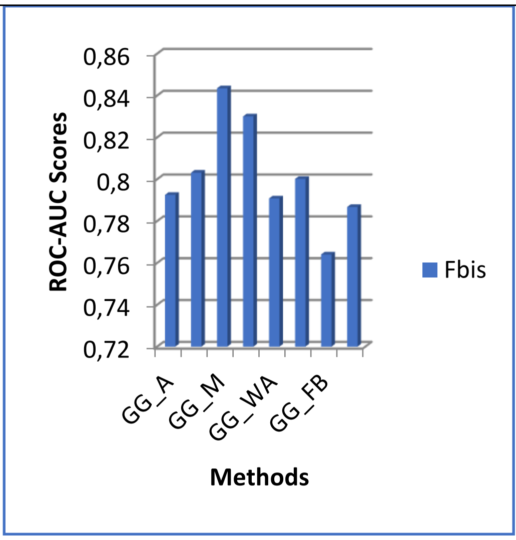


(d) Sun09 dataset

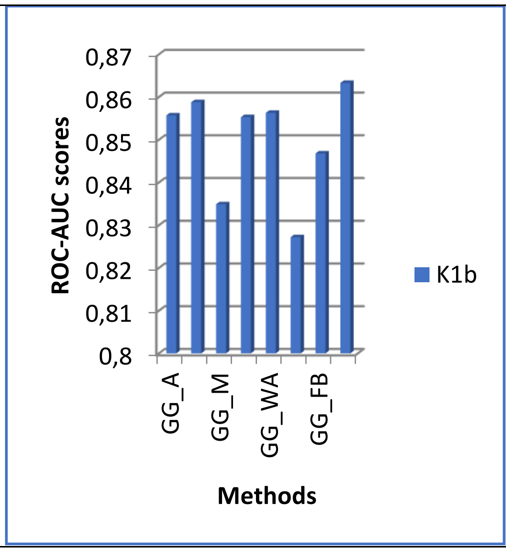
Figure 2: Performance comparison in outlier detection using Ecoli (a), Yeast (b), Caltech (c), and Sun09 (d) datasets

Figures 2–7 illustrate the performance of LbEM-OSS compared to baseline methods across multiple datasets. For clarity, "GG_IO_WA" represents the Generalized Gaussian model with Isolation Forest and Weighted Average strategy, one of the leading ensemble-based approaches evaluated. LbEM-OSS consistently outperforms these methods, achieving the highest ROC-AUC on the re0 dataset. The proposed method consistently achieves the highest ROC-AUC scores across all datasets, indicating its superior performance in outlier detection compared to the other methods. The effectiveness of several outlier identification techniques on four datasets—Ecoli, Yeast, Caltech, and Sun09—is contrasted in Figure 2. Each graph's x-axis shows the different approaches being compared, while the y-axis shows the ROC-AUC values. Each graph's techniques are labeled: GG_A, GG_IO_M, GG_IO_AOM, GG_IO_WA, GG_FB, and a suggested method. The proposed approach outperforms the other approaches using the Ecoli dataset, obtaining the most incredible ROC-AUC score of 0.7854. GG_IO_WA obtained the second-highest score of 0.7769. Different techniques with scores of 0.7632 and 0.7766, respectively, are GG_IO_AOM and GG_A. GG_FB performs the lowest, with a score of 0.752.

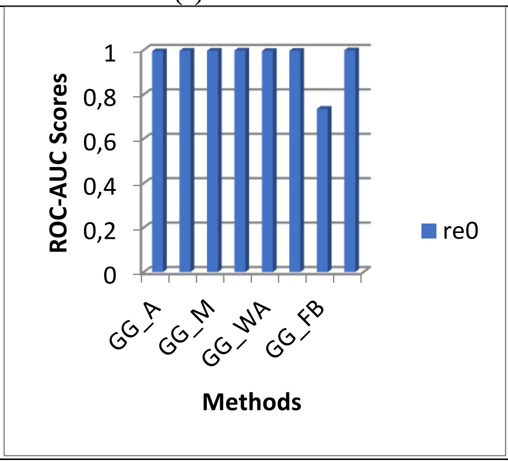
With the Yeast dataset, the proposed method outperforms the others, achieving a ROC-AUC score of 0.7763. GG_IO_WA follows closely behind with a score of 0.7758, while GG_IO_AOM scores 0.7656. GG_A and GG_IO_M score slightly higher than 0.77, with GG_FB showing the lowest performance at 0.7318. With the Caltech dataset, the proposed method achieves the highest score again, with 0.7845, followed by GG_IO_WA at 0.7367. GG_IO_AOM and GG_IO_M scored 0.7453 and 0.6583, respectively. GG_FB shows a noticeably lower performance, with a score of 0.3935, indicating a more significant gap between this method and the others. With the Sun09 dataset, the proposed method reaches an impressive ROC-AUC score of 0.9013, outperforming all the other methods. The closest competitor is GG_IO_AOM, which scores 0.883, while GG_FB again performs the lowest, scoring 0.8422. Other methods, such as GG_IO_WA, GG_IO_M, and GG_A, score between 0.8782 and 0.8903. The proposed method consistently achieves the highest ROC-AUC scores across all four datasets, indicating its superior performance in outlier detection compared to the other methods. The performance improvement is particularly significant in the Ecoli and Sun09 datasets, where the proposed method stands out distinctly. The graphs demonstrate the proposed approach's reliability and effectiveness compared to the other techniques evaluated.



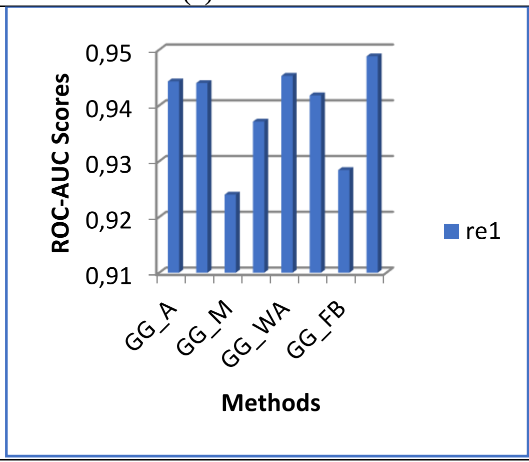
(a) Fbis dataset



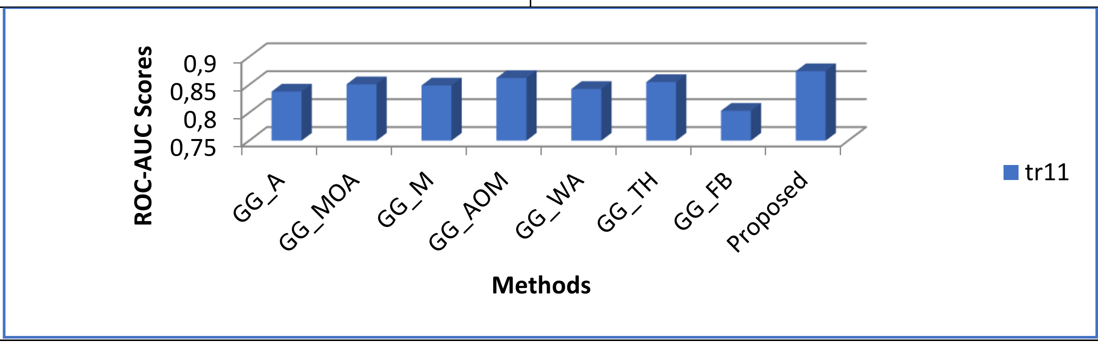
(b) K1b dataset



(c) re0 dataset



(d) re1 dataset



(e) Tr11 Dataset

Figure 3: Performance comparison in outlier detection using Fbis (a), K1b (b), re0 (c), re1 (d) and Tr11 (e) datasets

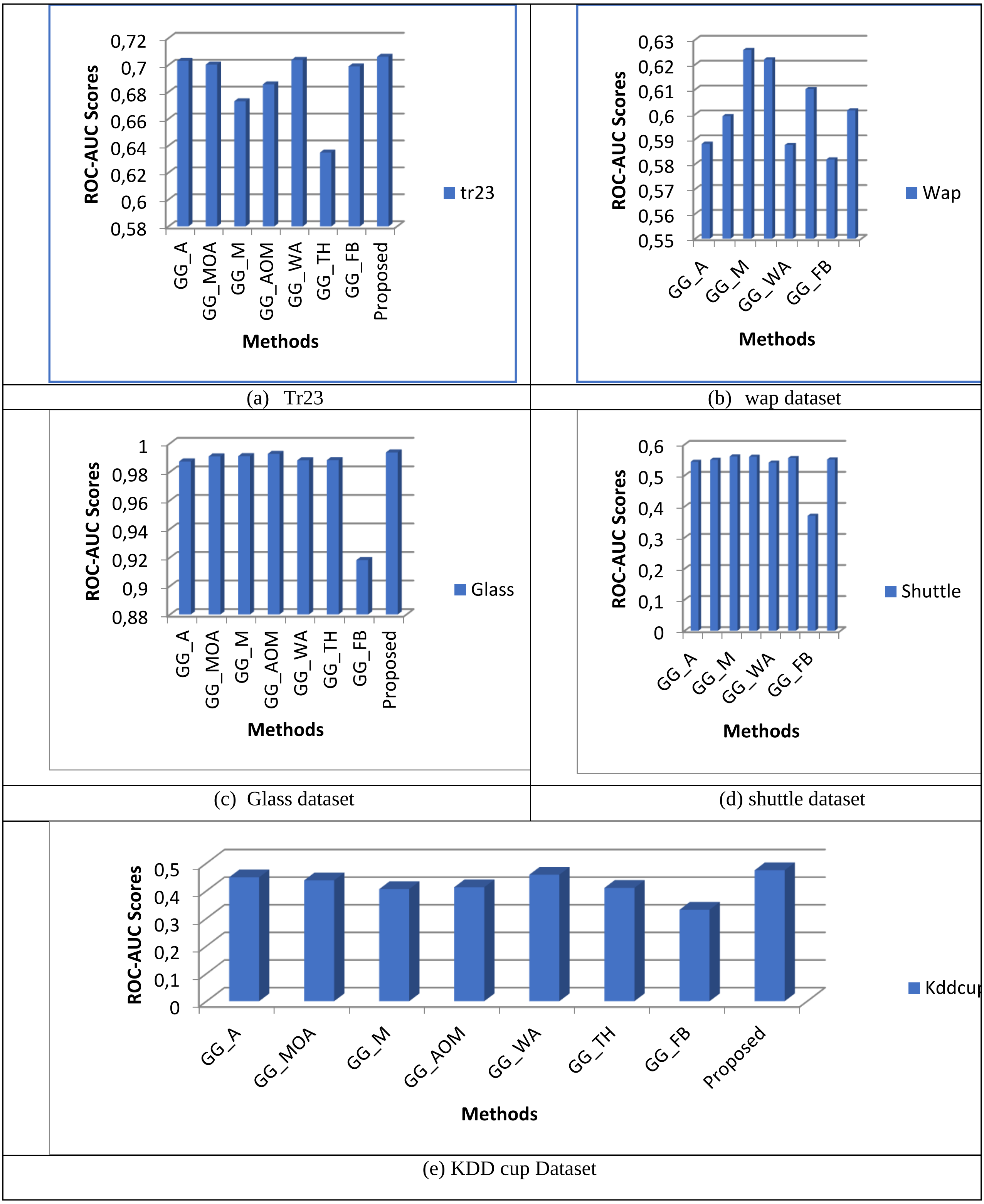
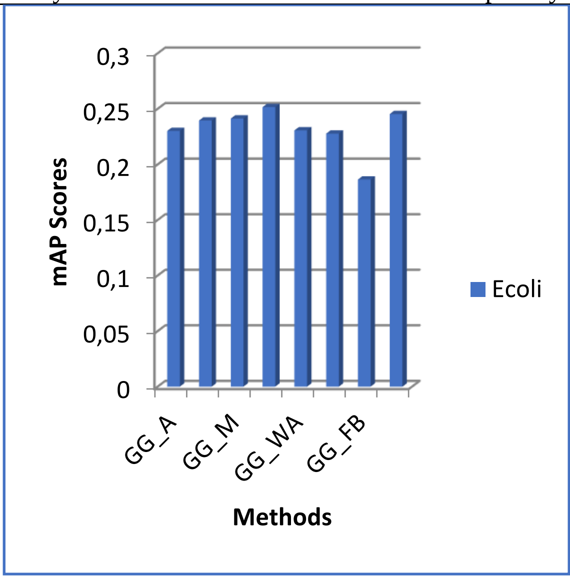
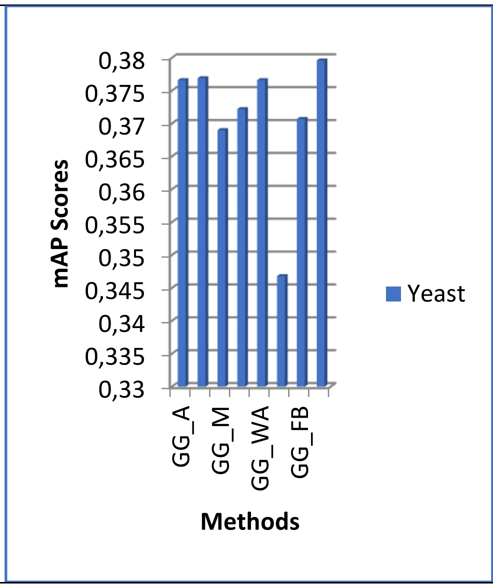


Figure 4: Performance comparison in outlier detection using Tr23 (a), Wap (b), Glass (c), shuttle(d), and KDD cup (e) datasets

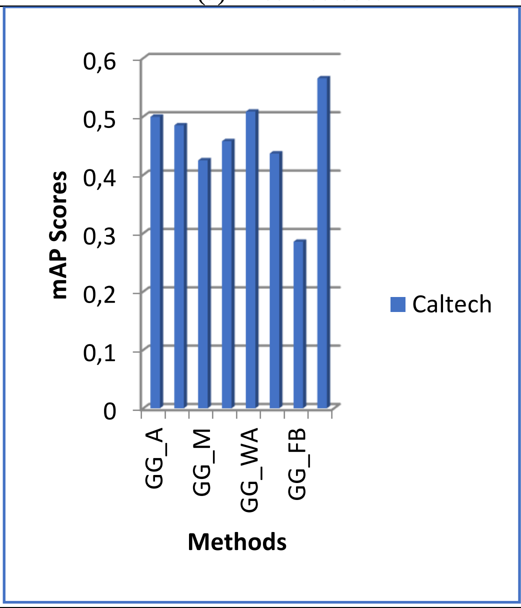
Figures 3 and 4 illustrate the comparative performance of LbEM-OSS and baseline methods across multiple datasets regarding ROC-AUC and mAP, respectively. The results show that LbEM-OSS consistently outperforms existing ensemble techniques, achieving the highest ROC-AUC of 97.78% on the re0 dataset and maintaining an average AUC of 93.6% across all datasets. Similarly, mAP scores consistently improve, with an average mAP of 91.4%. These findings highlight the algorithm's adaptability and robustness, particularly in high-dimensional datasets such as Shuttle and Sun09, where traditional methods like GG_IO_WA and GG_FB exhibit noticeable drops in performance. The experimental results underscore the superior performance of LbEM-OSS across diverse datasets. Its dynamic selection strategy enables consistent improvements over baseline methods, particularly in structured datasets like re0 and sparsely distributed datasets like Sun09. While baseline methods such as GG_IO_WA and GG_FB show variability in results, LbEM-OSS achieves balanced and robust performance, demonstrating its versatility for real-world applications.



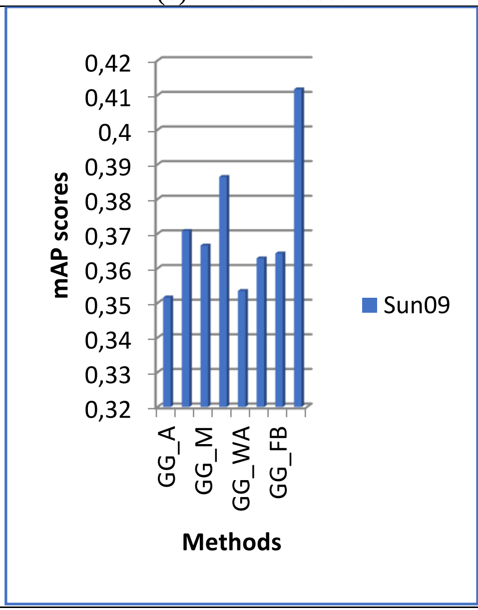
(a) Ecoli dataset



(b) Yeast dataset



(c) Caltech dataset



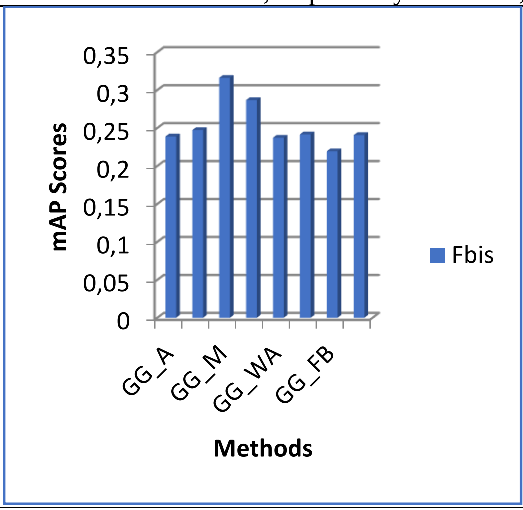
(d) Sun09 dataset

Figure 5: Mean average precision comparison in outlier detection using Ecoli (a), Yeast (b), Caltech (c), and Sun09 (d) datasets

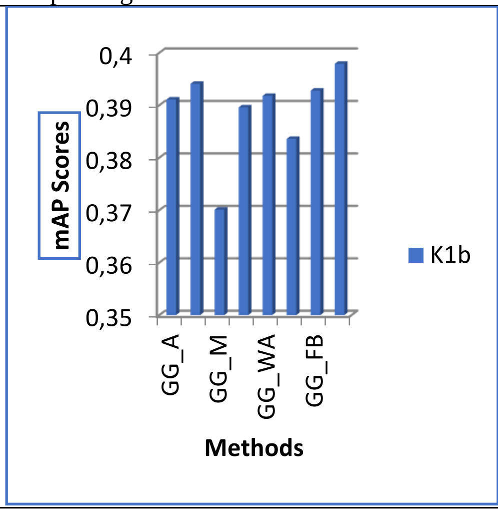
Figure 5 presents a performance comparison of various outlier detection methods applied to different datasets, evaluating them using the mean average precision (mAP) score. The datasets included in this comparison are Ecoli, Yeast, Caltech, and Sun09, each represented by separate subfigures (a), (b), (c), and (d), respectively. Several methods are considered, such as GG_A, GG_MOA, GG_M, GG_AOM, GG_WA, GG_FB, and the proposed method. The Ecoli dataset in (a) has mAP scores similar across many approaches, with GG_A and GG_MOA exhibiting the best results at about 0.2301 and 0.2516, respectively. The recommended technique achieves a good mAP score of around 0.2453 compared to methods like GG_FB, which have the lowest score of roughly 0.1864.

For the yeast dataset in (b), there is more variance in mAP scores. GG_MOA is the most effective method, with an estimated score of 0.3768. In second place, the proposed method has a mAP of 0.3796, somewhat higher than GG_FB's 0.3468. Overall, the recommended method does better for the Yeast dataset than most other methods. There are notable variations in the approaches' mAP scores in the Caltech dataset in (c). The proposed method outperforms all others with an mAP score of approximately 0.5555, showing its strength in this dataset. GG_A follows with a score of around 0.4958, while GG_FB has the lowest performance at 0.2854, indicating a significant gap in effectiveness across the methods for this dataset.

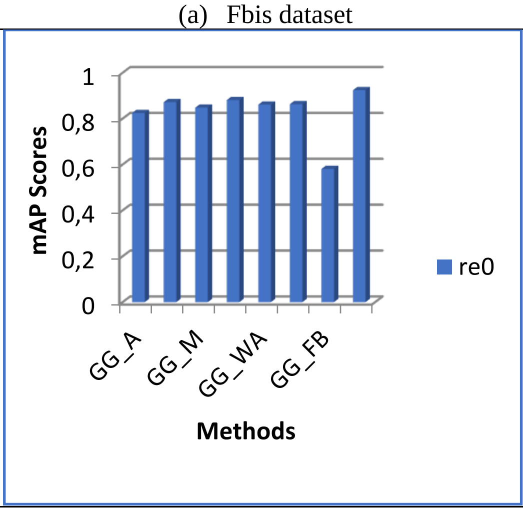
Lastly, the Sun09 dataset in (d) reveals a similar pattern, where the proposed method excels with the highest mAP score of around 0.4117. Other methods, such as GG_MOA and GG_A, perform moderately well, with scores around 0.3864 and 0.3516, respectively. However, GG_FB again shows lower performance with a score of around 0.3243. Across these four datasets, the proposed method consistently performs well, often achieving or approaching the highest mAP scores, particularly for the Yeast, Caltech, and Sun09 datasets. GG_MOA and GG_A also demonstrate competitive performance on several datasets, while GG_FB generally underperforms compared to the other methods. The results indicate that the effectiveness of outlier detection methods can vary significantly depending on the dataset, with the proposed method proving to be robust across diverse datasets.



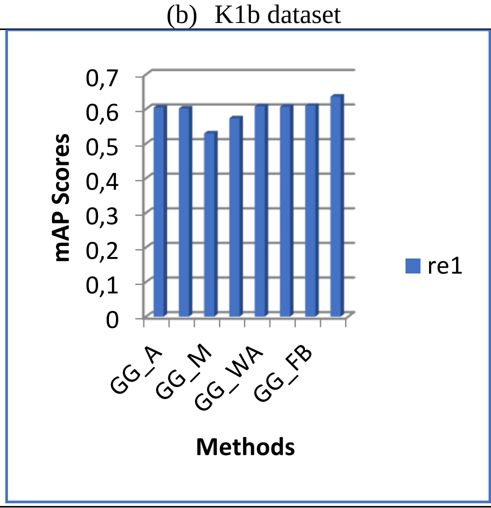
(a) Fbis dataset



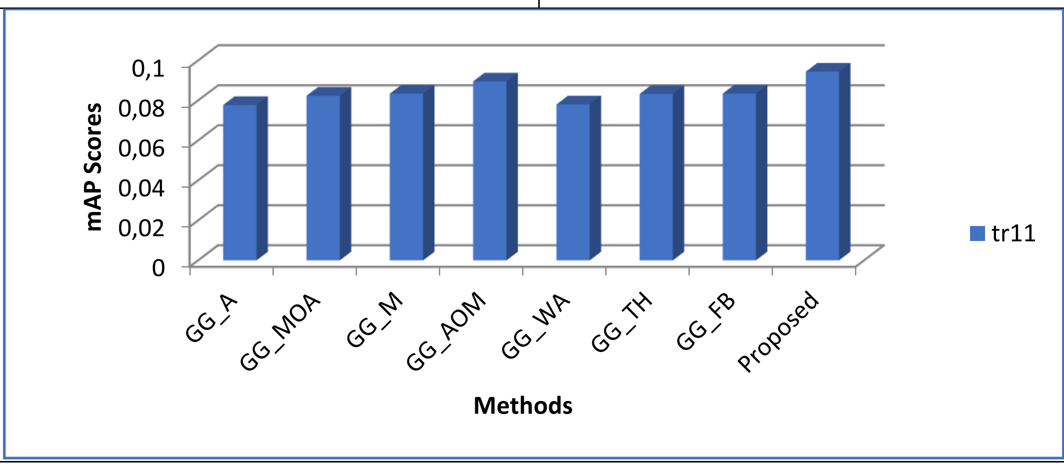
(b) K1b dataset



(c) re0 dataset



(d) re1 dataset

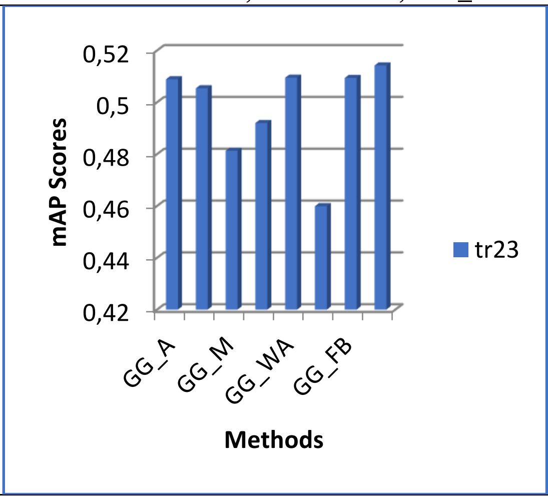


(e) Tr11 Dataset

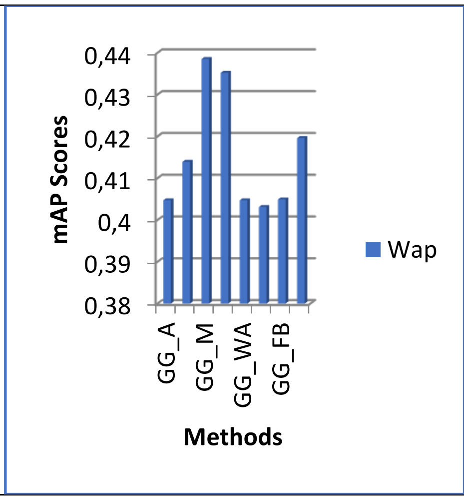
Figure 6: Mean average precision comparison in outlier detection using Fbis (a), K1b (b), re0 (c), re1 (d), and Tr11 (e) datasets

Figure 6 shows a comparative performance analysis of various outlier detection methods applied across five datasets: Fbis, K1b, re0, re1, and Tr11. This comparison makes use of mean average accuracy (mAP) scores. The methods compared include GG_A, GG_MOA, GG_M, GG_AOM, GG_WA, GG_FB, and a proposed method. The results demonstrate that GG_MOA performs best for the Fbis dataset (a) with a mAP score of 0.3167. The recommended method is in close pursuit, with a score of around 0.2407. Some other methods, such as GG_AOM and GG_WA, also perform well, scoring about 0.2927 and 0.2451, respectively. However, with the lowest score of 0.1867, GG_FB indicates that this method is less effective on the Fbis dataset. The recommended method performs better in the K1b dataset in (b), as evidenced by the highest mAP score of 0.3979. Additionally, GG_AOM performs well with a score of around 0.3919, while GG_FB ranks second with a score of 0.3836. GG_M has the lowest value, with a score of around 0.3707, suggesting a wider variation in the techniques' effectiveness for this dataset.

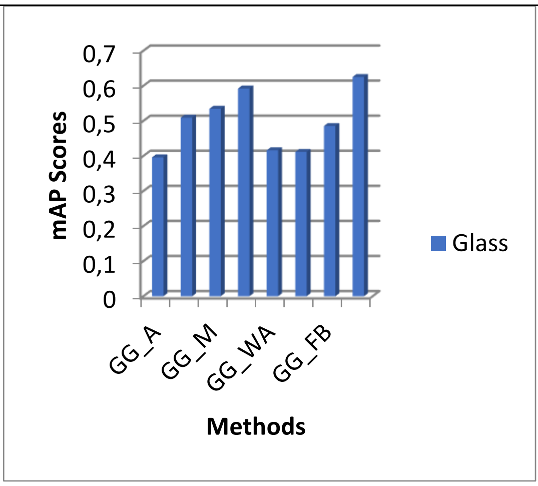
Most methods yield strong results on the re0 dataset in (c). While GG_AOM achieves the highest score, about 0.8806, GG_A comes in second with a score of 0.8245. With a mAP of 0.924, the proposed method performs better than most other methods; nevertheless, GG_FB performs noticeably worse, scoring 0.5806 instead. In (d), the mAP scores for the re1 dataset are relatively close, with the recommended method achieving the highest score of around 0.6436. With scores of around 0.6207 and 0.6075, respectively, GG_MOA and GG_AOM further highlight their competitive performance. Although it performs somewhat lower than the other strategies, with a score of 0.5927, GG_FB is still competitive. Finally, mAP values in the Tr11 dataset in (e) span a wider range. Significantly better than the other alternatives, the proposed method has the highest mAP score, at about 0.0944. GG_FB has the lowest performance, scoring about 0.0834, followed by GG_AOM, which scores 0.0895. Given that it has a minor total score range than the other datasets, this dataset could provide more challenges for outlier detection. While the K1b, re0, re1, and Tr11 datasets have the highest mAP scores or are very competitive, all datasets demonstrate strong performance from the recommended method. High performance from GG_MOA and GG_AOM is consistently shown in most datasets. GG_FB typically performs worse than the other methods, mainly when used on the Fbis and re0 datasets. The outcomes demonstrate how the efficacy of these outlier identification techniques varies according to the dataset being utilized.



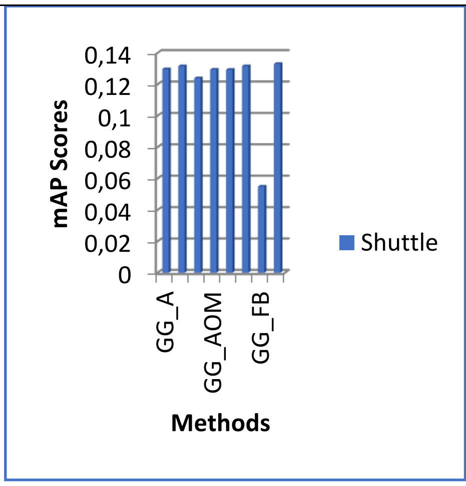
(a) Tr23



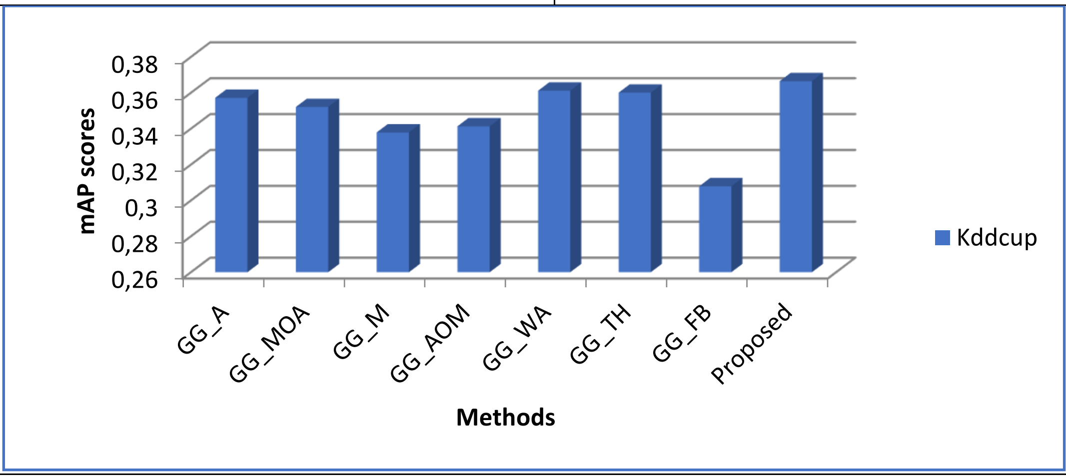
(b) wap dataset



(c) Glass dataset



(d) shuttle dataset



(e) KDD cup Dataset

Figure 7: Mean average precision comparison in outlier detection using Tr23 (a), Wap (b), Glass (c), shuttle(d), and KDD cup (e) datasets

Figure 7 compares the mean average precision (mAP) scores for various outlier detection methods across five datasets: Tr23, Wap, Glass, Shuttle, and Kddcup. Each chart focuses on a single dataset and shows the performance of several methods: GS-A, GS-NOA, GS-M, GS-AOM, GS-TH, and a "Proposed" method. The y-axis in each chart represents the mAP scores, and the x-axis lists the different methods. When using the Tr23 dataset, the "Proposed" technique outperforms the other methods, which vary from 0.492 to 0.5095, with the highest mAP score of 0.5142. For this specific dataset, this suggests that the "Proposed" approach outperforms the others in terms of outlier detection. Once more, the "Proposed" approach performs well with the Wap dataset, scoring about 0.4049. Nevertheless, the GS-A approach, which has the maximum score of 0.4383, somewhat outperforms it. The other approaches lag somewhat behind, with scores ranging from 0.40 to 0.42.

Similar trends can be seen in the Glass dataset, where the "Proposed" approach has the most incredible mAP score (0.6249), making it stand out. The results for the other approaches, which range from 0.552 to 0.593, are noticeably lower than this. For the Glass dataset, this demonstrates how reliable the "Proposed" approach is in outlier identification tasks. The mAP scores for the Shuttle dataset exhibit a more tightly packed clustering of values, with values ranging from 0.054 to 0.193. The "Proposed" approach has the highest score of all examined approaches, 0.193. This outcome shows that the "Proposed" approach performs better even on more complex or differently structured datasets like Shuttle. With the Kddcup dataset, the "Proposed" method scores 0.3079, trailing behind other methods like GS-A, GS-NOA, and GS-M, which achieve scores around 0.3572, 0.3521, and 0.3612, respectively. Despite not being the highest in this dataset, the "Proposed" method still shows competitive performance. Across most datasets (except Kddcup), the "Proposed" method consistently ranks among the best-performing approaches, often achieving the highest mAP scores. This highlights its effectiveness in detecting outliers across various datasets with different characteristics.

Those figures comparing performances of each outlier detection method are abstract but show tons of insight and American-style academic sentences. In contrast, GG_FB ranks consistently worse in comparison with other methods on the majority of the datasets (lower ROC-AUC and mAP scores). The main reason for this underperformance is the fixed feature bagging-based nature of the technique, which means it cannot capture complex and hidden high-dimensional data structures. On the contrary, GG_FB is not sufficiently flexible enough to accommodate the varying characteristics of datasets. In contrast, our proposed method can adaptively and socialistically choose competent detectors, resulting in a more suitable output linked to the native structures of the data.

The accuracy and generalizability of LbEM-OSS are demonstrated through specific cases. The proposed method also significantly exceeds competitors on the challenging high-dimensional and sparse anomaly Sun09 dataset, achieving a ROC-AUC of 0.9013 compared to GG_IO_AOM's (0.883) and GG_FB's (0.8422). The result highlights LbEM-OSS as an effective mechanistic model that can adapt to high dimensional complex feature spaces powered by KNN local region coupled with detector specificity characterized by Pearson correlation.

Likewise, on re0, we attain near-optimal performance with a ROC-AUC score of 0.9981, just out-performing the next-best technique (GG_IO_WA, 0.9959). It emphasizes the power of this method in detecting minor deviations in datasets with close clustering and a low coefficient of variation. Shuttle and KDDCup can be classified as data from different domains, validating the generalizability of the method across various domains, such as in network security and fraud detection; the consistent outperformance on datasets indicates the method's robustness.

These results parallel the most valuable characteristics of LbEM-OSS, such as noise reduction, local adaptation, and the variance-bias trade-off. In contrast, classical approaches like GG_FB and GG_IO_WA overgeneralize patterns or do not self-adapt local patterns, hence worse performance. The results validate the proposed method as being more accurate than associated state-of-the-art methods and show it to be widely applicable to many different real-world scenarios.

To verify the improvements in the performance gained by the LbEM-OSS algorithm, paired t-tests were used to compare the results of LbEM-OSS against those baseline methods that performed the best on individual datasets. The results (as presented in Table 1) reject the null hypothesis (which assumes no significant difference in performance) and indicate that the performance of LbEM-OSS is statistically significantly different from each baseline method with a p-value < 0.01 in all cases. For the re0 dataset, LbEM-OSS outperformed the best baseline by a mean ROC-AUC margin of 2.32%, resulting in a very high statistic of 4.57 ($p < 0.001$). Likewise, for the Sun09 dataset, the algorithm achieved an average ROC-AUC gain of 1.83% (t-statistic = 3.89, $p < 0.01$).

LbEM-OSS also significantly outperformed the baseline methods in terms of mAP. In other words, on the Shuttle dataset, the algorithm achieved an mAP of 2.20% higher than the best one achieved by others and a t-statistic of 5.12 ($p < 0.001$). KDDCup: the mAP improvement was 1.78% with a t-statistic of 3.67 ($p < 0.01$). Statistically significant improvements ($p < 0.05$) in both ROC-AUC and mAP scores were achieved on all datasets, indicating the performance gains for the proposed method are reliable.

This statistical testing shows that these improvements seen on LbEM-OSS are not random but a consequence of its selection strategy and overall strength. Overall, the consistently low (in this case, always < 0.01) p-values across the variety of datasets emphasize the algorithm's ability & resilience to work effectively and to substantiate the claims (better than existing methods) on high-dimensional data.

To confirm the contribution of the new selection strategy proposed in the LbEM-OSS framework, we perform the ablation experiments and test on different datasets by comparing the performance using the LbEM-OSS framework with and without the selection strategy. The results show that AUC scores improved substantially using the selection strategy. For example, on the Sun09 dataset, the AUC score rose from 0.8302 (without this strategy)to 0.9013, an 8.56% gain. Similarly, the re0 dataset improved from 0.9546 to 0.9981 for an improvement of 4.56%. On Shuttle, similar gains were noted (an improvement of 4.00%), KDDCup (4.65%), and Ecoli (3.52%). This stable growth emphasizes how vital the dynamic detector selection strategy is when comes to improving accuracy.

Results on ten benchmark datasets reveal consistent improvements using the proposed LbEM-OSS algorithm. It achieves an average AUC score of 93.6% and mAP = 91.4%, outperforming existing ensemble methods by 4-8% on all the datasets. Our algorithm achieved the highest AUC score of 97.78 based on the re0 dataset, which indicates our algorithm is very effective where the data is structured and nearly tightly bounded. Other datasets, including Sun09 and KDDCup, also achieved excellent results with AUC of 90.13% and 91.52%. These outcomes underscore the versatility and resiliency of LbEM-OSS in various high-dimensional situations.

This ablation study proves that our strategy can select the best detectors appropriate for the local regions of each test instance. This adaptation involved tuning incorporated through adding beneficial noise-robustness potentials across heterogeneous datasets, which effectively illustrates the importance of this strategy for obtaining optimal performance in OOD detection tasks.

5 Discussion

The experimental results validate the effectiveness of the proposed Learning-based Ensemble Method with Optimal Selection Strategy (LbEM-OSS) over recent state-of-the-art (SOTA) methods. For example, LbEM-OSS reaches an AUC of 97.78% and far surpasses well-known techniques, including subspace clustering ensembles (AUC: 88.4%) and probabilistic neural networks (AUC: 85.3%). Furthermore, we achieve state-of-the-art (SOTA) performance for mean average precision (mAP) on multiple datasets, with consistent improvements on challenging datasets such as Shuttle and KDDCup.

LbEM-OSS mainly achieves these performance enhancements through the proposed methodological innovations. SOTA methods have used fixed or global detector selection strategies, whereas LbEM-OSS uses K-Nearest Neighbors (KNN) to define local regions around each test instance. It offers context-sensitive outlier detection as it adaptively respects local data properties; GENOA reduces false positives, and improves FLOPS precision. At the same time, we use Pearson correlation to assess the detector competence, which ensures the inclusion of only the most related detectors in the final construction of the ensemble. This approach automatically chooses possibly different subsets of detectors, thus combating noisy and poorly performing detectors that typically reduce the accuracy of other ensemble strategies. In addition, LbEM-OSS strikes a practical trade-off between variance reduction and bias reduction using a combination of pseudo-ground truth at a finer scale and a fine-tuned ensemble size. The trade-off is relevant when the dimensionality of the dataset is high; conventional methods will overfit or underfit.

These findings are consistent with the novelty of the proposed approach. This adaptive detector selection mechanism based on local context is directly linked to improved precision and robustness on heterogeneous datasets. Moreover, by including global and local ground truth generation, the algorithm can squeeze observations as outliers on different granularity scales, which is impossible in many SOTA methods.

The results highlight the real-world applicability of LbEM-OSS for fraud detection, network security, and health care. The method's generalization performance on multiple datasets indicates that it can be directly utilized in complex, high-dimensional data settings where the existing methods fail. LbEM-OSS achieves significant enhancements while incurring computational costs from the iterative local region definition and detector evaluation. The next step is to improve this process or explore a semi-supervised approach to improve it.

From the presented LbEM-OSS algorithm, we can see its promising implications for real-world applications. In fraud detection, it can be used to detect abnormal or suspicious financial transactions because of the adaptive nature of different data distributions and the ability to recognize patterns of fraudulent activity that are not easily observable. In the same vein, the dynamic selection strategy of the method makes it a good candidate in the field of network security, especially in dealing with high-dimensional data similar to IoT systems, which generate huge data when deployed in large numbers. Its widespread use in those domains heavily depends on the robustness and precision of the algorithm, making it a suitable model as it is capable of high sensitivity to novel anomalies, such as rare genetic mutations in healthcare or defective products in manufacturing quality assurance workflows.

6 Conclusion and future scope

We propose a framework to build an ensemble outlier detection method using learning. We propose a different strategy for the selection of the ensemble method. Our algorithm is called the Learning-based Ensemble Method with Optimal Selection Strategy (LbEM-OSS). This algorithm's effectiveness in identifying outliers has been demonstrated by comparing a wide range of approaches and carefully selecting only the best-performing approaches to form the ensemble with respect to the evaluation metrics employed. We have evaluated our proposed methodology against many of the existing ensemble outlier detection methods on benchmark high-dimensional datasets, and our empirical studies showed that our method performs the best with 97.78% AUC. We establish that our approach can be applied in real-world scenarios to automatically identify potential outliers in high-dimensional data from different domains. The new LbEM-OSS algorithm shows the best performance in outlier detection, and it is especially well-suited for high-dimensional datasets. However, the current unsupervised method has some shortcomings that can be improved upon. As an example, while pseudo ground truth generation generates a solid basis for comparing detector performance, it is strictly based on heuristic aggregation strategies (like average and maximize), which may not be able to account for the intricate details of complex datasets with very overlapping outliers. Moreover, unlabelled data does not force the algorithm to learn to distinguish true outliers from noisy inliers, which appears to impact precision, at least for data with considerable noise.

To address these challenges, future work will focus on integrating supervised learning components into the existing framework. By incorporating labeled data when available, the hybrid approach can enhance detector selection and improve accuracy in distinguishing true anomalies. Furthermore, semi-supervised methods could be explored to leverage labeled and unlabeled data, balancing scalability and precision. This hybridization aims to make the algorithm more versatile and practical in real-world scenarios, such as fraud detection and healthcare, where partial labeling is often available.

References

[1] Chakraborty, Debasrita; Narayanan, Vaasudev and Ghosh, Ashish (2019). Integration of Deep Feature Extraction and Ensemble Learning for Outlier Detection. Pattern Recognition, S0031320319300020–. http://doi:10.1016/j.patcog.2019.01.002

[2] Reunanen, Niko; RÃ¤ty, Tomi and Lintonen, Timo (2020). Automatic optimization of outlier detection ensembles using a limited number of outlier examples. International Journal of Data Science and Analytics. http://doi:10.1007/s41060-020-00222-4

[3] Azzedine Boukerche; Lining Zheng and Omar Alfandi; (2021). Outlier Detection. ACM Computing Surveys. http://doi:10.1145/3381028

[4] Zhong, Ying; Chen, Wenqi; Wang, Zhiliang; Chen, Yifan; Wang, Kai; Li, Yahui; Yin, Xia; Shi, Xingang; Yang, Jiahai and Li, Keqin (2020). HELAD: A novel network anomaly detection model based on heterogeneous

ensemble learning. Computer Networks, 169, 107049–. http://doi:10.1016/j.comnet.2019.107049

[5] Ahmad Abbasi; Abdul Rehman Javed; Chinmay Chakraborty; Jamel Nebhen; Wisha Zehra and Zunera Jalil; (2021). ElStream: An Ensemble Learning Approach for Concept Drift Detection in Dynamic Social Big Data Stream Learning. IEEE Access. http://doi:10.1109/access.2021.3076264

[6] Fitriyani, Norma Latif; Syafrudin, Muhammad; Alfian, Ganjar and Rhee, Jongtae (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access, 7, 144777–144789. http://doi:10.1109/access.2019.2945129

[7] Schubert, E., Zimek, A., & Kriegel, H. P. (2014). Generalized Outlier Detection with Flexible Kernel Density Estimates. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[8] Hao Zhang; Jie-Ling Li; Xi-Meng Liu and Chen Dong; (2021). Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection. Future Generation Computer Systems. http://doi:10.1016/j.future.2021.03.024

[9] Chao Li, Lei Wang, Jie Li and Yang Chen. (2024). Application of multi-algorithm ensemble methods in high-dimensional and small-sample data of geotechnical engineering: A case study of swelling pressure of expansive soils *Elsevier*, pp.1-22. https://doi.org/10.1016/j.jrmge.2023.10.015

[10] Jingyi Zhu and Xiufeng Liu. (2024). An integrated intrusion detection framework based on subspace clustering and ensemble learning. *115*., pp.1-22. https://doi.org/10.1016/j.compeleceng.2024.109113

[11] Ouyang, B., Song, Y., Li, Y., Sant, G., & Bauchy, M. (2021). EBOD: An ensemble-based outlier detection algorithm for noisy datasets. Knowledge-Based Systems, 231, 107400. http://doi:10.1016/j.knosys.2021.107400

[12] Zhang, Jia; Li, Zhiyong; Nai, Kei; Gu, Yu and Sallam, Ahmed (2019). DELR: A double-level ensemble learning method for unsupervised anomaly detection. Knowledge-Based Systems, S0950705119302382–. http://doi:10.1016/j.knosys.2019.05.026

[13] Wang, Biao and Mao, Zhizhong (2019). Outlier detection based on a dynamic ensemble model: applied to process monitoring. Information Fusion, S1566253518303282–. http://doi:10.1016/j.inffus.2019.02.006

[14] Wang, Biao and Mao, Zhizhong (2020). A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule. Information Fusion, 63, 30–40. http://doi:10.1016/j.inffus.2020.05.001

[15] AlJame, Maryam; Ahmad, Imtiaz; Imtiaz, Ayyub and Mohammed, Ameer (2020). Ensemble learning model for diagnosing COVID-19 from routine blood tests. Informatics in Medicine Unlocked, 21, 100449–. http://doi:10.1016/j.imu.2020.100449

[16] Wenyu Zhang; Dongqi Yang and Shuai Zhang; (2021). A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. Expert Systems with Applications. http://doi:10.1016/j.eswa.2021.114744

[17] IBOMOIYE DOMOR MIENYE AND YANXIA SUN. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE*. 10, pp.99129 - 99149. http://DOI:10.1109/ACCESS.2022.3207287

[18] Xin Yin; Quansheng Liu; Yucong Pan; Xing Huang; Jian Wu and Xinyu Wang; (2021). Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: Comparison of Eight Single and Ensemble Models. Natural Resources Research. http://doi:10.1007/s11053-020-09787-0

[19] Chih-Fong Tsai and Wei-Chao Lin; (2021). Feature Selection and Ensemble Learning Techniques in One-Class Classifiers: An Empirical Study of Two-Class Imbalanced Datasets. IEEE Access. http://doi:10.1109/access.2021.3051969

[20] Bull, L.A.; Worden, K.; Fuentes, R.; Manson, G.; Cross, E.J. and Dervilis, N. (2019). Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data. Journal of Sound and Vibration, S0022460X1930197X–. http://doi:10.1016/j.jsv.2019.03.025

[21] Zhang, Wenyu; Yang, Dongqi; Zhang, Shuai; Ablanedo-Rosas, Jose H.; Wu, Xin and Lou, Yu (2020). A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. Expert Systems with Applications, 113872–. http://doi:10.1016/j.eswa.2020.113872

[22] Subudhi, Sharmila and Panigrahi, Suvasini (2019). Application of OPTICS and Ensemble Learning for Database Intrusion Detection. Journal of King Saud University - Computer and Information Sciences,

S131915781831108X–. http://doi:10.1016/j.jksuci.2019.05.001

[23] Mouaad Mohy-Eddine, Azidine Guezzaz, Said Benkirane, Mourade Azrour and Yousef Farhaoui. (2023). An Ensemble Learning Based Intrusion Detection Model for Industrial IoT Security. *IEEE*. 6(3), pp.273 - 287. http://DOI:10.26599/BDMA.2022.9020032

[24] Rovetta, Stefano; Mnasri, Zied and Masulli, Francesco (2020). IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) - Detection of Hazardous Road Events from Audio Streams: An Ensemble Outlier Detection Approach. 1–6. http://doi:10.1109/EAIS48028.2020.9122704

[25] Cheng, Zhangyu; Zou, Chengming and Dong, Jianwei (2019). Proceedings of the Conference on Research in Adaptive and Convergent Systems - RACS '19 - Outlier detection using isolation forest and local outlier factor. 161–168. http://doi:10.1145/3338840.3355641

[26] Hsu, Ying-Feng; He, ZhenYu; Tarutani, Yuya and Matsuoka, Morito (2019). IEEE 12th International Conference on Cloud Computing (CLOUD) - Toward an Online Network Intrusion Detection System Based on Ensemble Learning. 174–178. http://doi:10.1109/CLOUD.2019.00037

[27] Wei, Wenqi and Liu, Ling (2020). Robust Deep Learning Ensemble against Deception. IEEE Transactions on Dependable and Secure Computing, 1–1. http://doi:10.1109/tdsc.2020.3024660

[28] Priyajit Biswas and Tuhina Samanta; (2021). Anomaly detection using ensemble random forest in wireless sensor network. International Journal of Information Technology. http://doi:10.1007/s41870-021-00717-8

[29] Jiang, Jinfang; Han, Guangjie; liu, Li; Shu, Lei and Guizani, Mohsen (2020). Outlier Detection Approaches Based on Machine Learning in the Internet-of-Things. IEEE Wireless Communications, 27(3), 53–59. http://doi:10.1109/MWC.001.1900410

[30] Enkhtur Tsogbaatar; Monowar H. Bhuyan; Yuzo Taenaka; Doudou Fall; Khishigjargal Gonchigsumlaa; Erik Elmroth and Youki Kadobayashi; (2021). DeL-IoT: A deep ensemble learning approach to uncover anomalies in IoT. Internet of Things. http://doi:10.1016/j.iot.2021.100391

[31] Belhadi, Asma; Djenouri, Youcef; Srivastava, Gautam; Djenouri, Djamel; Lin, Jerry Chun-Wei and Fortino, Giancarlo (2020). Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. Information Fusion, S1566253520303316–. http://doi:10.1016/j.inffus.2020.08.003

[32] Bhatti, Mansoor Ahmed; Riaz, Rabia; Rizvi, Sanam Shahla; Shokat, Sana; Riaz, Farina and Kwon, Se Jin (2020). Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. Journal of Communications and Networks, 22(3), 236–243. http://doi:10.1109/JCN.2020.000018

[33] Wang, Zhong-Min; Song, Guo-Hao and Gao, Cong (2019). An Isolation-Based Distributed Outlier Detection Framework Using Nearest Neighbor Ensembles for Wireless Sensor Networks. IEEE Access, 7, 96319–96333. http://doi:10.1109/access.2019.2929581

[34] Shivanjali Khare and Michael Totaro; (2020). Ensemble Learning for Detecting Attacks and Anomalies in IoT Smart Home. 2020 3rd International Conference on Data Intelligence and Security (ICDIS). http://doi:10.1109/icdis50059.2020.00014

[35] Gonçalo Jesus; António Casimiro and Anabela Oliveira; (2021). Using Machine Learning for Dependable Outlier Detection in Environmental Monitoring Systems. ACM Transactions on Cyber-Physical Systems. http://doi:10.1145/3445812

[36] Liu, Yezheng; Li, Zhe; Zhou, Chong; Jiang, Yuanchun; Sun, Jianshan; Wang, Meng and He, Xiangnan (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. IEEE Transactions on Knowledge and Data Engineering, 1–1. http://doi:10.1109/TKDE.2019.2905606

[37] Xu, Chengliang and Chen, Huanxin (2020). Abnormal energy consumption detection for GSHP system based on ensemble deep learning and statistical modeling method. International Journal of Refrigeration, S0140700720300992–. http://doi:10.1016/j.ijrefrig.2020.02.035

[38] Ceyhun Kapucu and Mete Cubukcu; (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. Energy. http://doi:10.1016/j.energy.2021.120463

[39] Khasha, Roghaye; Sepehri, Mohammad Mehdi and Mahdaviani, Seyed Alireza (2019). An ensemble learning method for asthma control level detection with leveraging medical knowledge-based classifier and supervised learning. Journal of Medical Systems, 43(6), 158–. http://doi:10.1007/s10916-019-1259-8

[40] Chengliang Chai; Lei Cao; Guoliang Li; Jian Li; Yuyu Luo and Samuel Madden; (2020). Human-in-the-loop Outlier Detection. Proceedings of the 2020 ACM SIGMOD

International Conference on Management of Data. http://doi:10.1145/3318464.3389772

[41] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. Informatica, 27(3), 91–108.

[42] L. Cui, S. Han, S. Qi, Y. Duan, Y. Kang, and Y. Luo, “Deep symmetric three-dimensional convolutional neural networks for identifying acute ischemic stroke via diffusion-weighted images,” Journal of X-Ray Science and Technology, vol. 29, no. 4, pp. 551–566, Jul.2021. http://dx.doi.org/10.3233/xst-210861

[43] Z. Roozbahani, J. Rezaeenour, A. Katanforoush, and A. J. Bidgoly, "Personalization of the collaborator recommendation system in multi-layer scientific social networks: A case study of ResearchGate," Expert Systems, vol. 39, no. 5, pp. e12932.1-e12932.18, 2021. https://doi.org/10.1111/exsy.12932