

Benchmarking Outlier Detection Algorithms in Healthcare: An Analysis Across Diverse Medical Datasets

Jaloliddin Rustamov

700043175@uaeu.ac.ae

United Arab Emirates University

Zahiriddin Rustamov

United Arab Emirates University

Nadia Badawi

United Arab Emirates University

Rafat Damseh

United Arab Emirates University

Fatma Al Jasmi

United Arab Emirates University

Nazar Zaki

United Arab Emirates University

Article

Keywords:

Posted Date: November 19th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-5242162/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Benchmarking Outlier Detection Algorithms in Healthcare: An Analysis Across Diverse Medical Datasets

Jaloliddin Rustamov^{1,2,3,*}, Zahiriddin Rustamov^{1,4}, Nadia Badawi¹, Rafat Damseh¹, Fatma Al Jasmi², and Nazar Zaki^{1,*}

¹Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, 15551, Abu Dhabi, United Arab Emirates

²Department of Genetics and Genomics, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, 15551, Abu Dhabi, United Arab Emirates

³Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, 3000, Belgium

⁴Department of Computer Science, KU Leuven, Leuven, 3000, Belgium

*700043175@uaeu.ac.ae;nzaki@uaeu.ac.ae

ABSTRACT

Outlier detection plays a vital role in data analysis by identifying data points that differ from the expected patterns. In healthcare, this is especially important, as spotting these anomalies can help detect unusual patient conditions, errors in data, or even early signs of fraud. However, even though this is crucial, little research has compared how well different outlier detection methods work in healthcare. This study aims to fill that gap by testing 37 different outlier detection algorithms on 27 healthcare datasets, covering everything from common conditions like diabetes to rare diseases like thyroid disorders. We measured their performance using metrics like ROC-AUC, Average Precision, and F1 Score, along with practical considerations like training time, prediction speed, and memory usage. Our results show that no single model works best across all scenarios. However, methods like HBOS, COPOD, and IForest achieved average ROC-AUC scores of 0.85, 0.83, and 0.82, respectively, performing well in various situations. More complex models like AnoGAN and DeepSVDD achieved ROC-AUC scores up to 0.90 on specific datasets, particularly excelling in datasets with higher dimensionality. This variety highlights the need to choose outlier detection methods based on the specific healthcare context, balancing accuracy and efficiency to meet different clinical and practical needs. This paper provides a solid foundation for future research and development of outlier detection systems in healthcare.

Introduction

Outliers, also called anomalies, are data points that stand out, often indicating something unusual. Outlier detection algorithms aim to find these deviations, which could reveal important patterns or potential problems^{1,2}. Outlier detection, a critical data analysis component, is particularly vital in healthcare settings. In medical contexts, outliers often represent anomalous patient conditions, data entry errors, or early indicators of critical health issues. Accurately identifying these anomalies can significantly impact patient care, from early disease detection to refinement of treatment protocols.

Healthcare data presents unique challenges for outlier detection algorithms. The high dimensionality of medical records, which may include imaging data, genomic sequences, and detailed patient histories, can hinder the performance of traditional algorithms due to the curse of dimensionality³. Class imbalance is prevalent, as rare diseases constitute a small fraction of the population, making it difficult for models to learn from insufficient positive cases⁴. Moreover, the critical impact of errors in healthcare settings is profound; false positives can lead to unnecessary treatments and patient anxiety, while false negatives may result in missed diagnoses and delayed interventions⁵.

Outlier detection methods can be broadly categorized into three types, each suited for different scenarios in healthcare. Supervised outlier detection uses labeled datasets to train a classifier, similar to traditional pattern recognition. However, this approach faces challenges due to the inherent imbalance in medical data, where outliers are typically far less common than normal cases. Semi-supervised outlier detection utilizes datasets containing only normal data to create a model of typical behavior, with outliers detected as deviations from this pattern. This approach can be particularly useful in scenarios where anomalies are rare or difficult to collect. Unsupervised outlier detection, the most flexible approach, doesn't require labeled data. Instead, it analyzes the intrinsic characteristics of the data, using metrics like distance or density to identify outliers^{1,6}.

While numerous outlier detection methods exist, their effectiveness in healthcare settings remains understudied, particularly

with tabular medical data across various diseases. This research aims to comprehensively evaluate existing outlier detection techniques and architectures in a healthcare context, focusing on their ability to handle diverse tabular medical datasets. Our study encompasses a wide range of conditions, including but not limited to appendicitis, various cancers, metabolic disorders, neurodegenerative diseases, and post-surgical outcomes. We assess these methods based on their accuracy, robustness to class imbalance, interpretability, and computational efficiency.

By identifying the most reliable techniques for detecting rare diseases or abnormal results in medical data, this research seeks to provide healthcare professionals with improved tools for early diagnosis and intervention. Furthermore, our findings will contribute to developing more effective outlier detection strategies tailored to the unique demands of medical data analysis. This study compares several advanced techniques to identify the most reliable ones for healthcare, particularly for detecting rare diseases or abnormal test results. Our evaluation will help determine which methods deliver high accuracy and address the unique challenges posed by medical data, including the critical balance between sensitivity and specificity in clinical decision-making.

Review of the Related Work

Several studies have evaluated outlier detection algorithms across different datasets, providing useful insights into their performance, reliability, and computational efficiency.

Gaspar et al.⁷ looked at techniques for detecting outliers in healthcare data. However, their work was limited to a review and lacked critical or empirical analysis. Since then, new methods have been developed, but more research specifically focused on healthcare data is still needed. Most previous studies have been domain-agnostic, overlooking the unique challenges of healthcare data, which highlights the need for a more focused comparison in this area.

Campos et al.² compared unsupervised outlier detection methods on various datasets, including some from healthcare. Their study examined how different algorithms performed and how parameters affected results. They found that traditional methods like k -Nearest Neighbors (k NN), k -Nearest Neighbors Weighted (k NNW), and Local Outlier Factor (LOF) were still strong performers, especially for datasets with larger outliers. However, the study noted that the selection of methods was not exhaustive and that class downsampling may have introduced some bias.

Goldstein and Uchida¹ compared 19 unsupervised outlier detection algorithms across ten public datasets from various fields, including two from healthcare. They examined detection performance, computation time, and how parameter settings affected results. They found that nearest-neighbor-based methods performed better than clustering-based methods, with k NN excelling for global tasks and LOF for local tasks. However, their study didn't include healthcare-specific insights.

Emmott et al.⁸ conducted a meta-analysis of outlier detection algorithms, focusing on factors like dataset characteristics and problem dimensions. They created benchmarks by adjusting four variables: point difficulty, the frequency of anomalies, how clustered the anomalies were, and feature relevance. Isolation Forest performed best, particularly with high-dimensional data, but the study didn't focus on healthcare datasets and had some limitations related to benchmark difficulty.

Domingues et al.⁹ studied the performance, robustness, and efficiency of outlier detection algorithms, along with memory usage. They found that Isolation Forest, Robust Kernel Density Estimation, Probabilistic Principal Component Analysis, and One-Class SVM performed well. However, they didn't offer specific insights into healthcare applications.

Xu et al.⁶ compared nine popular outlier detection algorithms across 18 datasets, including seven from healthcare. They aimed to provide a detailed overview of outlier detection methods, focusing on their performance with high-dimensional and mixed-type data. They found that neighbor-based methods were more reliable, while ensemble methods showed inconsistent results. The study didn't go into detail about undersampling methods for positive cases and lacked a focus on healthcare-specific applications.

Steinbuss and Böhm¹⁰ developed a process to generate realistic synthetic data for benchmarking outlier detection methods. They used real-world data to create normal data and outliers for better comparisons. They found that different methods worked best for various types of outliers, with no single method being the best for all cases. The study had some limitations, including possible bias and the need for better approaches for specific types of outliers.

Soenen et al.¹¹ studied how hyperparameter tuning affects the performance of outlier detection algorithms, using 16 datasets, including seven from healthcare. They found that hyperparameter selection greatly influenced the results and suggested that a small labeled validation set could help tune parameters more effectively. However, since their focus wasn't specifically on healthcare, the findings may not fully apply to this field.

Han et al.¹² compared 30 outlier detection algorithms across 57 datasets, examining their performance with varying levels of supervision, different anomaly types, and noisy data. They found no unsupervised algorithm consistently outperformed the others, but semi-supervised methods did better with just 1% labeled anomalies. However, their study did not cover structured data types like time series, limiting its relevance to some real-world applications.

Building on these studies, our research critically evaluates outlier detection methods using healthcare datasets. We focus on the unique challenges of healthcare data, offering insights and recommendations on which algorithms are best suited for

Table 1. Past studies on comparative evaluation of outlier detection algorithms

Paper	Datasets	Healthcare Datasets	Positive Class Undersampling Method (Outlier Ratio)	Algorithms Tested	Evaluation Metrics
²	23	11	Random (up to 5%)	12	ROC-AUC, $\overline{PR}$
¹	10	2	First 10 positive cases	19	ROC-AUC
⁸	19	1	Random (up to 1%)	8	ROC-AUC, $\overline{PR}$
⁹	15	2	Random (0.20%-7%)	14	ROC-AUC, $\overline{PR}$
⁶	18	7	—	9	ROC-AUC, R-Precision
¹¹	16	7	Random (up to 18.7%)	6	Average AUC
¹⁰	19	9	Same as ²	12	ROC-AUC, $\overline{PR}$
¹²	57	12	Random (up to 40%)	30	ROC-AUC, $\overline{PR}$
Ours	27	27	Random stratified (1%) or min. 3 obs.	37	ROC-AUC, $\overline{PR}$, $F1$, SE , SP

$\overline{PR}$ = Avg Precision; $F1$ = F1 Score; SE = Sensitivity; SP = Specificity

different applications. Our study expands on previous research by using a broader set of healthcare datasets and considering factors like dataset characteristics and the potential of ensemble and semi-supervised methods. Table 1 provides an overview of datasets and methods used in past studies.

Methodology

Datasets

In this study, we compiled a diverse collection of 27 healthcare datasets ranging from 4 features up to 452 features and 72 samples up to 2111 samples from various sources, including the UCI Machine Learning Repository¹³, KEEL¹⁴, and other research papers or studies. These datasets encompass a wide range of diseases and purposes, such as appendicitis, breast and cervical cancer, diabetes, obesity, cardiovascular diseases, Parkinson’s disease, thyroid disorders, post-surgical survival rates, hematological diseases, orthopedic conditions, liver cirrhosis, lymphography, Alzheimer’s disease, hepatitis, and stroke. Specifically, we included the UCTH breast cancer data¹⁵, a cardiovascular disease dataset by Al-Shamsi et al.¹⁶, and the PIMA diabetes dataset¹⁷. This diverse collection of datasets aims to provide a comprehensive testbed for critically evaluating the performance of various outlier detection methods in the healthcare domain.

Data Preprocessing

In this section, we discuss the data preprocessing steps we performed. If the dataset was not addressed, it did not require data preprocessing. Details about the dimensions of the datasets before and after preprocessing have been outlined in Table 2. The target class was renamed to *Class* for uniformity across all datasets. This standardization facilitates the application of programmatic reductions, simulating scenarios where positive classes are treated as outliers. Then, the categorical variables within the dataset were identified and transformed using a one-hot encoding method from the scikit-learn library¹⁸. This transformation is crucial for preparing categorical data for the ML models, which typically require numerical input. By converting categorical variables into a binary matrix representation, one-hot encoding eliminates any ordinal relationships that might be misinterpreted by the models, ensuring that each category is treated as a separate entity without any implicit order or magnitude. This process aids in improving model accuracy and interpretability. The one-hot encoder is configured to ignore any unknown categories encountered in the dataset, enhancing the model’s robustness and flexibility when dealing with unseen data. All other non-categorical features in the dataset are retained in their original form. Past studies have indicated that outliers typically have a relative frequency ranging from 1% (0.01) to 10% (0.1) in comparison to normal data points¹⁹. Another study claims that observations are considered outliers when their proportion is lower than 5%⁹. Given the lack of a standard defining outlier ratio and the variability observed in past studies as shown in Table 1, we have set the outlier ratio at 1%, with a minimum of three observations in cases where 1% of the dataset comprises fewer than three observations. The details of datasets after reduction and preprocessing are presented in Table 3. The detailed preprocessing performed on each dataset and its code is available at <https://github.com/jaloliddin-rus/outlier-detection-healthcare.git>

Architectures Employed

Probabilistic Models for Outlier Detection

The probabilistic outlier detection models explored in this study offer diverse approaches to estimating the likelihood of data points being anomalies based on probabilistic measures and density estimations. The Stochastic Outlier Selection (SOS)²⁰ converts dissimilarities into affinities to compute outlier probabilities, adapting variance with a perplexity parameter to enhance

Table 2. Dimension of datasets before and (after) preprocessing

Dataset	Samples (After)	Features (After)	Positives (After)	Negatives (After)	Outlier Ratio
Appendicitis	106 (106)	8 (8)	21 (21)	85 (85)	0.1981
Breast_cancer	213 (205)	11 (10)	93 (89)	120 (116)	0.4341
CVD_UAE	977 (977)	19 (18)	97 (97)	880 (880)	0.0993
Haberman	306 (306)	4 (4)	81 (81)	225 (225)	0.2647
HCC_data	165 (165)	50 (50)	102 (102)	63 (63)	0.6182
Heart	303 (303)	14 (14)	165 (165)	138 (138)	0.5446
Mammographic	962 (830)	6 (6)	445 (403)	516 (427)	0.4855
Pima	768 (768)	9 (9)	268 (268)	500 (500)	0.3490
Spectfheart	267 (267)	45 (45)	212 (212)	55 (55)	0.7940
Stroke	749 (749)	11 (11)	249 (249)	500 (500)	0.3324
Wisconsin	699 (683)	10 (10)	241 (239)	458 (444)	0.3499
Heart_failure	299 (299)	13 (13)	96 (96)	203 (203)	0.3211
Obesity	2111 (2111)	17 (17)	1824 (1824)	287 (287)	0.8640
Hepatitis	155 (155)	20 (18)	32 (32)	123 (123)	0.2065
Parkinsons	195 (195)	24 (23)	147 (147)	48 (48)	0.7538
Arrhythmia	430 (402)	279 (279)	185 (165)	245 (237)	0.4104
Ilpd	583 (579)	11 (11)	416 (414)	167 (165)	0.7150
Darwin	174 (174)	452 (451)	89 (89)	85 (85)	0.5115
Lymphography	148 (148)	19 (19)	2 (2)	146 (146)	0.0135
Breast_cancer_coimbra	116 (116)	10 (10)	64 (64)	52 (52)	0.5517
Cirrhosis	418 (418)	19 (17)	161 (161)	257 (257)	0.3852
Vertebral	310 (310)	7 (7)	210 (210)	100 (100)	0.6774
Cervical_cancer	72 (72)	20 (20)	21 (21)	51 (51)	0.2917
Diabetic_retinopathy	1151 (1151)	20 (20)	611 (611)	540 (540)	0.5308
Bone_marrow	187 (182)	37 (37)	85 (81)	102 (101)	0.4451
Thoracic_surgery	470 (470)	17 (17)	70 (70)	400 (400)	0.1489
Thyroid	215 (215)	6 (6)	65 (65)	150 (150)	0.3023

sensitivity to local data structures. Similarly addressing data distribution uniformity, the Wrap-around Quasi-Monte Carlo Discrepancy (WQMCD)²¹ assesses how well sample points fill the parameter space, with higher discrepancy scores indicating potential anomalies. Kernel Density Estimation (KDE)²² and Gaussian Mixture Models (GMM)²³ both estimate data density to detect outliers; KDE does so non-parametrically by summing kernel functions assigned to each data point, while GMM assumes data are generated from a mixture of Gaussian distributions and uses the Expectation-Maximization algorithm to estimate parameters, identifying outliers based on low membership probabilities. Diverging from distance-based methods, Angle-Based Outlier Detection (ABOD)²⁴ utilizes angles and variances at each point to remain effective in high-dimensional spaces, with small angles suggesting potential outliers. Both Copula-Based Outlier Detection (COPOD)²⁵ and Empirical-Cumulative-distribution-based Outlier Detection (ECOD)²⁶ employ empirical cumulative distribution functions to estimate tail probabilities, where low tail probabilities signal anomalies. Lastly, the Rapid Distance-Based Outlier Detection via Sampling²⁷ leverages sampling strategies and distance measures, computing the distance to the nearest neighbor within a sample subset to quantify outlierness. Collectively, these models provide varied probabilistic frameworks for outlier detection, differing in their assumptions about data distribution, computational complexity, and adaptability to different data structures.

Linear Models and Dimensionality Reduction Techniques

The linear outlier detection models evaluated in this study utilize statistical techniques and transformations to identify anomalies based on linear relationships within the data. Cook's Distance (CD)²⁸ measures the influence of each observation in linear regression by assessing how its removal affects the estimated regression coefficients, flagging data points with large CD values as potential outliers due to their significant impact. Similarly focusing on data influence, the Linear Method for Deviation Detection (LMDD)²⁹ detects deviations by evaluating how excluding subsets of data reduces overall dissimilarity, efficiently identifying outliers in large datasets through a smoothing factor without relying on predefined error patterns. The Minimum Covariance Determinant (MCD)³⁰ robustly estimates the central tendency by finding the subset of observations with the smallest covariance determinant, detecting outliers based on their leverage and being effective against data contamination. Both Principal Component Analysis (PCA)³¹ and Kernel PCA (KPCA)³² perform dimensionality reduction to capture significant patterns; PCA does so linearly to facilitate outlier detection as anomalies deviate from principal components, while KPCA extends this to nonlinear data distributions using kernel methods, measuring novelty through reconstruction error. The One-Class Support Vector Machine (OCSVM)³³ establishes a hyperplane in high-dimensional space to encompass normal data, isolating outliers by optimizing the margin and allowing a fraction of data points outside the boundary, similar to KPCA's use of kernel

Table 3. Dimension of datasets after preprocessing and reduction

Dataset	Samples	Feature Count	Positives	Negatives	Outlier Ratio
Appendicitis	88	8	3	85	0.0341
Breast_cancer	119	10	3	116	0.0252
CVD_UAE	889	18	9	880	0.0101
Haberman	228	4	3	225	0.0132
HCC_data	66	50	3	63	0.0455
Heart	141	14	3	138	0.0213
Mammographic	435	6	8	427	0.0184
Pima	507	9	7	500	0.0138
Spectfheart	58	45	3	55	0.0517
Stroke	507	11	7	500	0.0138
Wisconsin	450	10	6	444	0.0133
Heart_failure	206	13	3	203	0.0146
Obesity	308	17	21	287	0.0682
Hepatitis	126	18	3	123	0.0238
Parkinsons	51	23	3	48	0.0588
Arrhythmia	241	279	4	237	0.0166
Ilpd	170	11	5	165	0.0294
Darwin	88	451	3	85	0.0341
Lymphography	148	19	2	146	0.0135
Breast_cancer_coimbra	55	10	3	52	0.0545
Cirrhosis	261	17	4	257	0.0153
Vertebral	103	7	3	100	0.0291
Cervical_cancer	54	20	3	51	0.0556
Diabetic_retinopathy	551	20	11	540	0.0200
Bone_marrow	104	37	3	101	0.0288
Thoracic_surgery	404	17	4	400	0.0099
Thyroid	153	6	3	150	0.0196

functions for adaptability. Collectively, these linear models leverage statistical properties, dimensionality reduction, and support vector mechanisms to detect outliers, differing in their handling of data structures, computational complexity, and normality assumptions.

Distance and Density-Based Methods

The proximity-based outlier detection models investigated in this study employ techniques centered around distances, densities, and connectivity among data points to identify anomalies. The Rotation-based Outlier Detection (ROD)³⁴ is a parameter-free method that detects outliers by decomposing data into three-dimensional subspaces and applying geometric rotations using Rodrigues' rotation formula to analyze volume discrepancies with the Median Absolute Deviation (MAD), making it adaptable to complex data structures without assuming any statistical distribution. The k -nearest-neighbor (kNN) algorithm³⁵ identifies global anomalies by computing distances to the k nearest neighbors, with anomaly scores based on either the distance to the k^{th} neighbor or the average distance to all k neighbors; however, its effectiveness hinges on the appropriate choice of k and it does not detect local anomalies. The Cluster-Based Local Outlier Factor (CBLOF)³⁶ utilizes clustering algorithms like k -means to calculate anomaly scores that consider both the distance to cluster centers and cluster sizes. However, relying on cluster size as a weighting factor has been questioned, and the initial number of clusters significantly influences results. Histogram-based Outlier Score (HBOS)³⁷ constructs individual histograms for each feature under the assumption of feature independence, calculating outlier scores by multiplying inverse bin heights; this method offers computational efficiency, especially in high-dimensional datasets but may overlook feature dependencies. The Local Outlier Factor (LOF)³⁸ detects local anomalies by comparing the local density of a data point to that of its neighbors, assigning higher LOF scores to points with significantly lower densities, making it practical for identifying local outliers but sensitive to the choice of k . Addressing parameter selection challenges, the Local Correlation Integral (LOCI) algorithm³⁹ considers all possible values of k by expanding an r -neighborhood and comparing densities of differently sized neighborhoods controlled by a parameter α , though at the cost of increased computational complexity ($O(n^3)$). The Connectivity-based Outlier Factor (COF)⁴⁰ modifies LOF by incorporating the chaining distance to account for the connectedness of data points, effectively detecting outliers in datasets with linear dependencies or non-spherical distributions by constructing a minimum spanning tree rooted at each point. Lastly, the Subspace Outlier Detection (SOD) algorithm⁴¹ targets high-dimensional data by identifying low-variance subspaces shared among a point and its neighbors, calculating outlier scores based on deviations within these subspaces, which allows for effective detection of outliers in datasets with varying correlations. These proximity-based models offer diverse approaches to outlier detection by leveraging distance metrics, density estimations, and connectivity analyses, each with unique considerations

regarding computational efficiency, parameter sensitivity, and applicability to different data structures.

Ensemble Methods for Outlier Detection

The outlier ensemble models evaluated in this study combine multiple detection techniques to enhance anomaly identification by leveraging the strengths of individual methods. The Deep Isolation Forest (DIF)⁴² enhances the traditional Isolation Forest approach by integrating deep learning techniques to address limitations such as difficulty in detecting complex anomalies in high-dimensional and non-linearly separable data spaces. DIF employs a novel representation scheme utilizing casually initialized neural networks to map original data into random representation ensembles. This mapping allows the model to perform random axis-parallel cuts on these new representations, effectively translating to non-linear partitions in the original data space. Similarly, the Isolation using Nearest Neighbor Ensemble (iNNE)⁴³ leverages the nearest neighbor ensemble to isolate anomalies by dividing the dataset into regions using a subsample and calculating an isolation score for each region. By adapting to the local distribution of the data, iNNE generates isolation scores that serve as local measures referring to the local neighborhood, enabling the detection of global and local anomalies. Feature Bagging⁴⁴, designed for high-dimensional outlier detection, operates by creating an ensemble of base detectors that each analyze random subspaces of the feature set. This method selects a random number of features, typically between half and nearly all the total features available, to form varied sub-datasets in each iteration. An outlier detection algorithm, such as LOF (Local Outlier Factor), is applied to these subsets. LOF's scoring system, which is inherently normalized, alleviates the need for further normalization across different subspace sizes, simplifying the process. Scores from these multiple detectors are aggregated through a breadth-first ranking or a cumulative sum of scores to establish a final outlier score. LODA (Lightweight On-line Detector of Anomalies)⁴⁵ is another ensemble-based outlier detection method that combines multiple weak outlier detectors to achieve robust performance. It generates a series of random projections of the input data into one-dimensional spaces, each involving only a subset of the original features. For each projection, LODA constructs a histogram to estimate the probability density of the data points. The final outlier scores are calculated as the mean negative log-likelihood of each data point according to the histograms of the random projections. XGBOD (Extreme Gradient Boosting Outlier Detection)⁴⁶ is a hybrid outlier detection method that combines unsupervised outlier detection algorithms with a supervised classifier. It begins by applying multiple unsupervised outlier detectors to the input data, which generate new features capturing different aspects of the data's outlier behavior. These extracted features are then concatenated with the original features, creating an augmented feature space. Finally, an XGBoost classifier is trained to identify anomalies in this enhanced feature space. XGBOD aims to improve the overall performance of outlier detection by leveraging the strengths of both unsupervised and supervised techniques. These outlier ensemble models offer diverse approaches to enhancing outlier detection by integrating multiple detection methods. Methods like DIF and iNNE extend traditional isolation-based approaches to better capture complex anomalies by utilizing deep learning or adapting to local data distributions. Feature Bagging and LODA leverage random subspaces and projections to create robust ensembles against irrelevant attributes and efficiently combine multiple weaker detectors to reveal outliers effectively. XGBOD uniquely integrates unsupervised and supervised techniques, augmenting the feature space to improve anomaly detection. Each model has unique considerations regarding computational efficiency, adaptability to different data structures, and the ability to detect complex or local anomalies, offering versatile tools for robust outlier detection in various data contexts.

Neural Network-Based Approaches

The neural network-based outlier detection models explored in this study employ various deep learning techniques to capture complex data patterns and identify anomalies. Deep Support Vector Data Description (DeepSVDD)⁴⁷ utilizes a neural network to learn a compact representation of normal data by minimizing the volume of a hypersphere that encompasses the network representations in the latent space, effectively detecting anomalies that deviate from this learned normal representation. Similarly focusing on reconstruction errors, the AutoEncoder model²³ comprises an encoder and decoder that compress and reconstruct data, with significant reconstruction errors indicating potential outliers due to deviation from typical data distributions. Extending this concept probabilistically, the Variational Auto-Encoder (VAE)⁴⁸ employs a probabilistic framework with stochastic variational inference, optimizing a lower bound estimator of the likelihood function; outliers are identified based on high reconstruction errors within this probabilistic model. Adopting adversarial frameworks, the Adversarially Learned Anomaly Detection (ALAD)⁴⁹ model utilizes a bi-directional Generative Adversarial Network (GAN) with both a generator and an encoder, leveraging cycle consistency and reconstruction errors to flag anomalies, enhancing efficiency and accuracy in high-dimensional data environments. The Single-Objective Generative Adversarial Active Learning (SO-GAAL)⁵⁰ employs a generative adversarial framework where a generator creates synthetic data points that mimic potential outliers. A discriminator differentiates these from normal data, aiming to generate data points that overlap with actual outliers. Addressing mode collapse in GANs, the Multi-Objective Generative Adversarial Active Learning (MO-GAAL)⁵⁰ extends SO-GAAL by incorporating multiple generators with distinct objectives, producing a diverse distribution of synthetic data points to better define the boundary between outliers and normal observations. Lastly, the AnoGAN model⁵¹ employs a deep convolutional GAN to learn the manifold of normal data, mapping query images to the closest point in the latent space through an iterative process, with

the anomaly score derived from the mapping loss effectively identifying anomalies based on deviation from learned normal appearance. Collectively, these neural network models utilize deep learning architectures—including autoencoders, variational inference, and adversarial training—to capture complex data structures and identify anomalies, each offering unique approaches to handling high-dimensional data, reconstruction errors, and the generation of synthetic data to enhance outlier detection.

Graph-based Approaches

The graph-based outlier detection models explored in this study leverage graph structures and neural network techniques to identify anomalies based on the relationships and connectivity among data points. The Learnable Unified Neighbourhood-based Anomaly Ranking (LUNAR) model⁵² utilizes Graph Neural Networks (GNNs) to enhance local outlier detection methods like LOF and DBSCAN by integrating their principles into a unified GNN framework. This approach treats data points as nodes in a graph with edges representing neighborhood relationships, allowing LUNAR to learn from nearest neighbors' information via a trainable system that not only detects but also ranks anomalies based on their deviation from expected neighborhood patterns, thereby achieving robust outlier detection across varied datasets with superior flexibility and performance. Complementarily, the R-Graph model⁵³ addresses outlier detection within datasets approximating a union of low-dimensional subspaces by utilizing data self-representation and random walks on a graph. Building on the self-expressiveness property, where each data point within a subspace can be represented as a linear combination of other points in the same subspace, R-Graph constructs an asymmetric affinity matrix and forms a weighted directed graph. Random walks initiated from data points on this graph reveal outliers by exploiting connectivity patterns; data points that do not return to their initial position are identified as outliers. The model's effectiveness relies on the geometric arrangement and connectivity of data points within their respective subspaces, offering theoretical guarantees under specific data distribution and graph connectivity assumptions. Collectively, these graph-based models exploit the structural and relational information inherent in data graphs, employing GNNs and graph traversal methods to detect anomalies, each providing unique advantages in handling complex data structures and connectivity patterns for robust outlier detection.

Model Training

In this study, we employed a train-test split, which differs from the standard way of applying this method, where we first separated the dataset into positive and negative cases based on the 'Class' variable. Then, we randomly sampled a subset of negative cases, equal to 20% of the total negative cases, to form the test set. This test set was combined with all the positive cases and shuffled to ensure randomness. The remaining negative cases were used as the training set and shuffled. We implemented one-hot encoding using the scikit-learn library to encode categorical variables. We used the Python toolbox for scalable outlier detection PyOD⁵⁴ to implement the above methods discussed in earlier sections. All the experiments were performed on a desktop computer with specifications: Ryzen 9 5950X, 128GB RAM, RTX 3090 24GB.

Evaluation Metrics

With 37 algorithms and 27 datasets, we have 999 micro experiments that need to be evaluated. We cannot use a common performance metric, such as accuracy, as true negatives will introduce bias to the result. Therefore, we employed two of the most common metrics in the past literature: ROC-AUC and average precision. We also utilized F1-Score, sensitivity, and specificity to evaluate the predictive performance of outlier detection models assessed in this study.

Results

This section presents the comprehensive findings from our evaluation of 37 outlier detection algorithms applied across 28 healthcare datasets. Each model was assessed based on its performance in identifying outliers within these highly specialized datasets, which vary significantly in features and sample sizes. Our analysis focuses on several key performance metrics that offer insights into the effectiveness and efficiency of these models in a healthcare context. We employed ROC-AUC and average precision scores to evaluate the detection capabilities of each algorithm. Additionally, F1-Score, sensitivity, and specificity were used to gauge the models' ability to accurately distinguish outliers from normal observations. Beyond these traditional performance metrics, we also examined the computational cost of each model by measuring the time taken for training and prediction, as well as the memory usage during these processes. These factors are crucial in practical applications, where accuracy and operational efficiency are paramount. The results are organized into two main sections. First, we provide a detailed analysis of each algorithm's performance across the various datasets, highlighting specific strengths and weaknesses regarding detection accuracy and computational demands. Second, we offer a comparative overview that ranks the algorithms, facilitating a clear understanding of which methods are most suitable for specific types of healthcare data based on their performance and resource utilization. By presenting these results, we aim to equip researchers and practitioners with the necessary information to select appropriate outlier detection techniques for their specific needs in the healthcare domain. This section reflects on the algorithms' capability to identify rare, significant medical conditions and their practical applicability in real-world settings

where computational resources and response times are often limited. The models that encountered errors during training on certain datasets outlined in Table 1 in supplementary materials were excluded from evaluation alongside other models, where we compared mean performance values such as ROC-AUC, F1 Score, and average precision.

Figure 1 provides a detailed analysis of the performance of various outlier detection models across our dataset collection, as measured by the Mean ROC-AUC and Mean Average Precision.

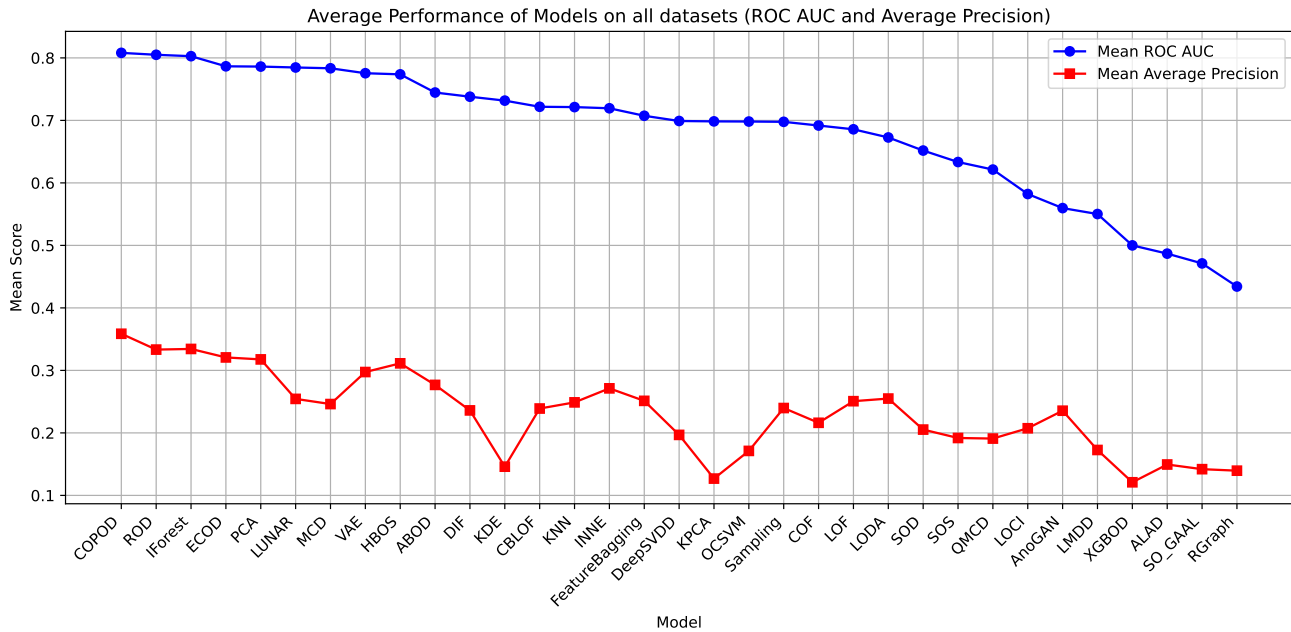


Figure 1. Average performance of models on datasets (Mean ROC-AUC and average precision)

Illustrated by the blue line, the ROC-AUC scores consistently remain above 0.6 for most models at the start of the graph, suggesting a solid capability in ranking the outliers. High ROC-AUC values indicate that these models effectively distinguish between outlier and normal observations, which is crucial for reliable outlier detection in medical applications. The red line, representing average precision scores, shows more significant variability and generally lower scores than ROC-AUC. This metric, reflecting the precision-recall balance, highlights the challenges in maintaining high precision, which is particularly important in healthcare settings where the cost of false positives is high. Models such as COPOD, ROD, and ECOD start with notably higher scores in both metrics, demonstrating their robustness and effectiveness. These models manage to correctly rank outliers and maintain a higher true positive rate, which is essential for practical applications like early disease detection or identifying erroneous entries in medical records. Notably, as we progress towards models such as DeepSVDD, AnoGAN, and RGraph on the graph, there is a marked decrease in both ROC-AUC and average precision scores. This trend could suggest that while complex models might be theoretically appealing, their practical application requires careful calibration to avoid overfitting and ensure they generalize well across varied healthcare datasets. The divergence in performance between the ROC-AUC and average precision scores for specific models emphasizes the need to select outlier detection algorithms based on their ability to rank outliers and their precision in real-world settings. This selection is crucial in healthcare, where the implications of false positives and negatives can be significant.

We extend our analysis to encompass the F1 Score, Sensitivity, and Specificity, depicted in Figure 2. These metrics provide a deeper understanding of each model's performance, highlighting their ability to balance detecting true positives with avoiding false positives—critical considerations in healthcare.

The F1 Score, represented by the blue line, combines precision and recall into a single metric, offering a gauge of a model's overall efficacy in handling the delicate balance between identifying true positive cases and minimizing false positives. Across the models, we observe a generally stable F1 Score, suggesting that many models manage this balance reasonably well. However, none excel across all datasets, indicating potential overfitting or underfitting issues in specific contexts. Sensitivity, shown by the green line, measures the true positive rate, which is paramount in medical applications to ensure that all potential outlier cases, such as diseases, are captured. The graph reveals significant variability in sensitivity among the models. High peaks suggest that specific models like KPCA can capture true positive cases. Still, these peaks are often followed by sharp declines in specificity, reflecting inconsistencies in model performance. Specificity, illustrated by the red line, assesses each

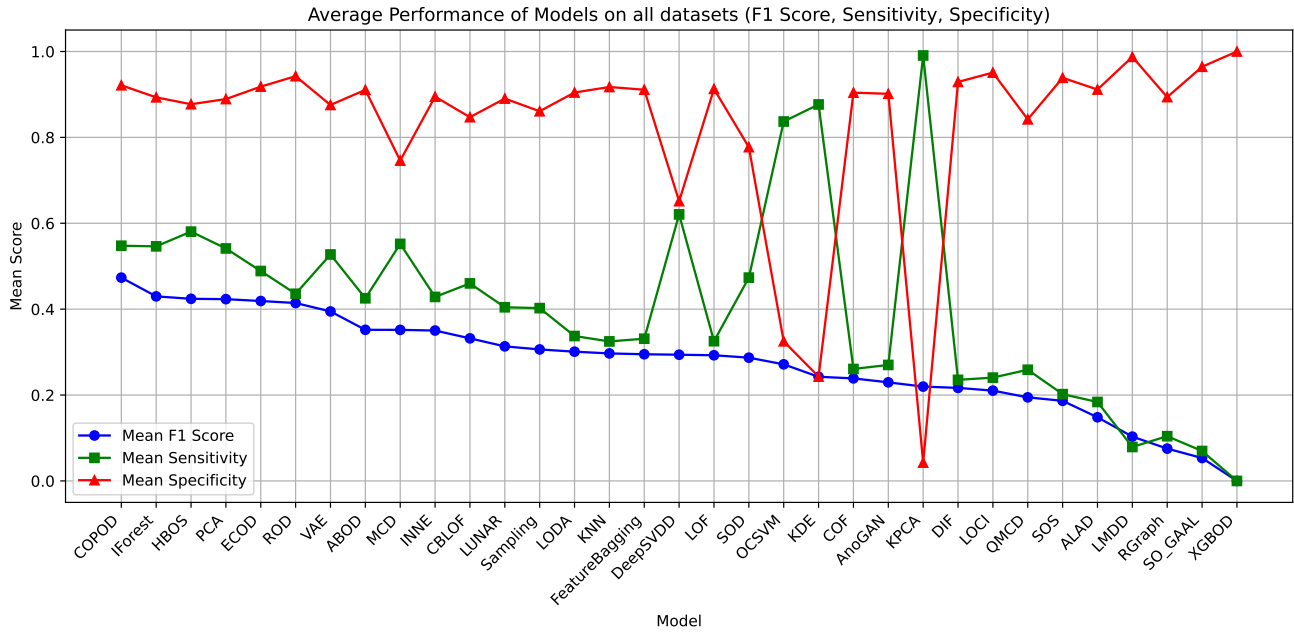


Figure 2. Average performance of models on datasets (Mean F1 Score, sensitivity and specificity)

model's ability to identify true negatives correctly. High specificity is crucial in healthcare settings to prevent false alarms, which can lead to unnecessary diagnostic procedures and patient anxiety. The specificity scores are notably high for many models, particularly those toward the end of the graph. However, this often comes at the expense of sensitivity, illustrating a trade-off between avoiding false positives and the risk of missing true positives. The combined analysis of these three metrics illustrates the varied performance of outlier detection models across different scenarios. Models like ROD and ECOD display a commendable balance across all three metrics, suggesting their potential suitability for a wide range of healthcare applications. Conversely, some models exhibit excellent sensitivity but falter in F1 Score and specificity, which could limit their utility in scenarios where precision is paramount.

Figures 3, 4 and 5 provide an insightful perspective by showcasing the performance of the models (in terms of Average Precision, F1 Score and ROC-AUC) sorted by the upper quartile. This analysis highlights how these models perform at least 75% of the time, offering a robust measure of their effectiveness in a clinical setting.

Following our exploration of F1 Score, Sensitivity, and Specificity, we delve deeper into the average precision of different models, as presented in Figure 3. This box plot representation provides a statistical summary emphasizing the median, variability, and the presence of outliers in the precision scores across various datasets. The box plots in Figure 3 reveal a wide range of average precision values, highlighting the variability in model performance across different healthcare datasets. Models such as COPOD, IForest, and HBOS exhibit higher median precision values with narrower interquartile ranges, suggesting a relatively consistent performance across various datasets. This consistency is crucial for applications in healthcare, where the reliability of outlier detection can significantly impact clinical decisions and outcomes. Conversely, models towards the right of the figure, such as LODA, AnoGAN, and XGBOD, show a broader range of precision scores, evidenced by longer whiskers and the presence of outliers (depicted by circles). These models exhibit high variability, which might stem from their sensitivity to dataset specifics, such as the number of features, noise, or the imbalance in class distributions.

Expanding upon our detailed analysis of average precision, we next examine the F1 scores for different outlier detection models, as shown in Figure 4. The box plots in Figure 4 display a broad range of F1 scores across the models, revealing both median performance and variability within each model. Notably, models such as ROD, COPOD, and IForest at the left end of the graph exhibit relatively higher median F1 scores with tighter interquartile ranges, suggesting consistent and reliable performance across various datasets. This reliability is essential for applications where the consequences of false diagnostics can be critical, such as in disease diagnosis and patient monitoring. In contrast, models towards the right of the graph, such as AnoGAN, SOS, and XGBOD, demonstrate lower median F1 scores and more significant variability, as indicated by longer whiskers and the presence of outliers. The extensive range in these scores indicates potential over-sensitivity to specific dataset characteristics, which may lead to inconsistent performance.

Continuing our examination of model performance metrics, we now focus on the ROC-AUC scores, as shown in Figure 5.

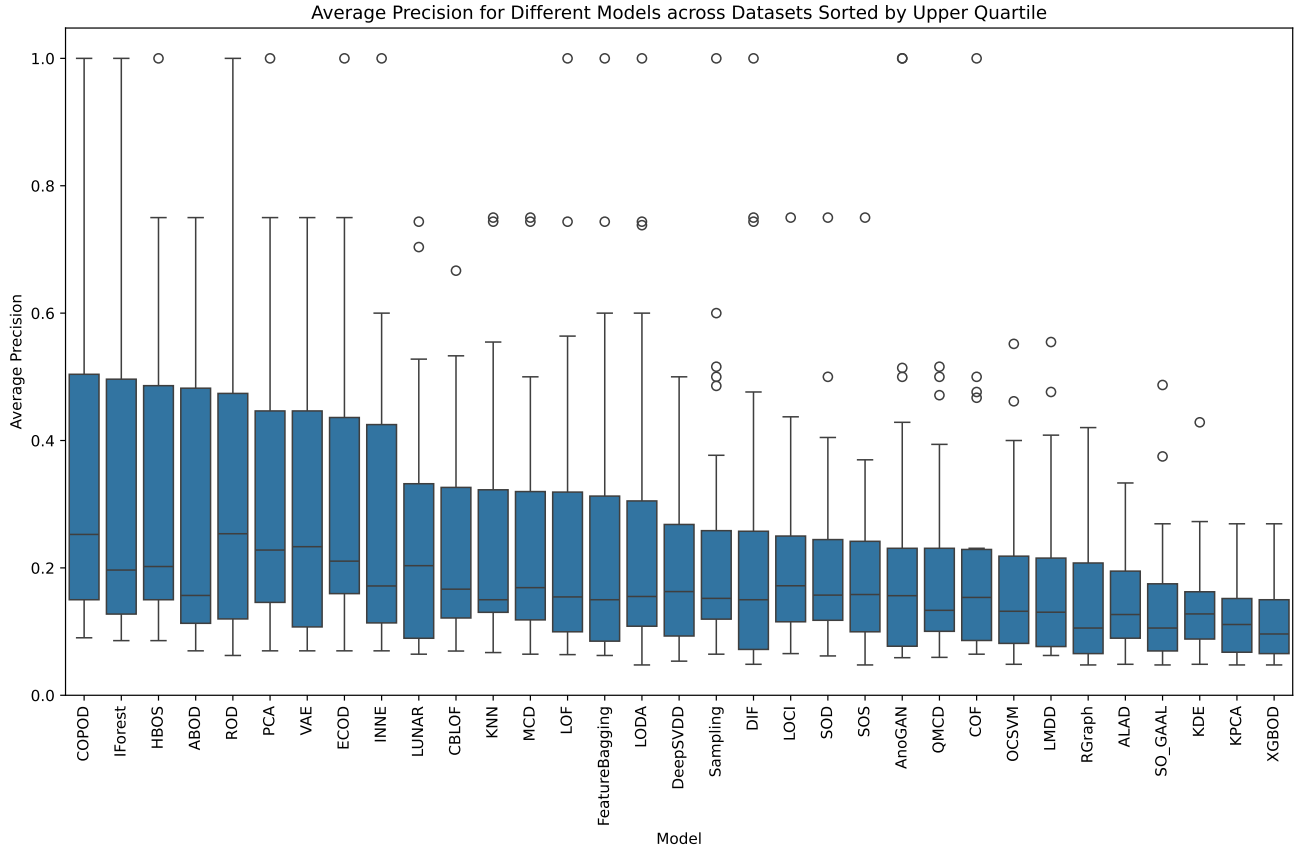


Figure 3. Average Precision for Different Models across Datasets Sorted by Upper Quartile

Figure 5 presents a box plot distribution of mean ROC-AUC scores for various models, illustrating the median performance and variability within each model across different datasets. The plot shows that models such as IForest, HBOS, and COPOD, which are positioned on the left, typically exhibit higher median ROC-AUC scores with relatively narrow interquartile ranges, which suggests that these models consistently identify true positives and negatives across various healthcare datasets, making them reliable choices for critical applications such as disease detection and patient monitoring. Conversely, models towards the right end of the plot, such as LODA, SOS, and XGBOD, display lower median scores and greater spread in their interquartile ranges, highlighted by the presence of outliers and extended whiskers, which may indicate that they did not perform well in some datasets.

Continuing our comprehensive evaluation of outlier detection models, we shift our focus to operational metrics, explicitly examining the average training and prediction times, along with memory usage for each model, as illustrated in Figure 6. These metrics are critical for practical applications in healthcare, where computational resources and response times can significantly influence the feasibility and efficiency of deploying these models in real-time settings. The graphs depict the computational demands associated with each model. The first graph shows that specific models, such as ABOD and AnoGAN, exhibit significantly higher training times, which could be limiting in scenarios where models must be retrained frequently or on enormous datasets. Similarly, the prediction time graph highlights that some models like RGraph, though quick to train, may take longer to make predictions, which could impede their use in real-time applications. Memory usage, both during training and prediction, as shown in the third and fourth graphs, further compounds the operational considerations. Models that require substantial memory resources may not be suitable for deployment in environments with hardware limitations or where multiple applications must run simultaneously. These operational metrics and the earlier performance evaluations provide a holistic view of each model's practicality in healthcare contexts. For instance, while some models may score highly on accuracy metrics like ROC-AUC or F1 Score, their utility may be compromised by high computational demands. Conversely, models demonstrating moderate performance but lower computational requirements might be more viable for specific healthcare applications, especially in resource-constrained settings.

Following our extensive review of individual performance and operational metrics of outlier detection models, we com-

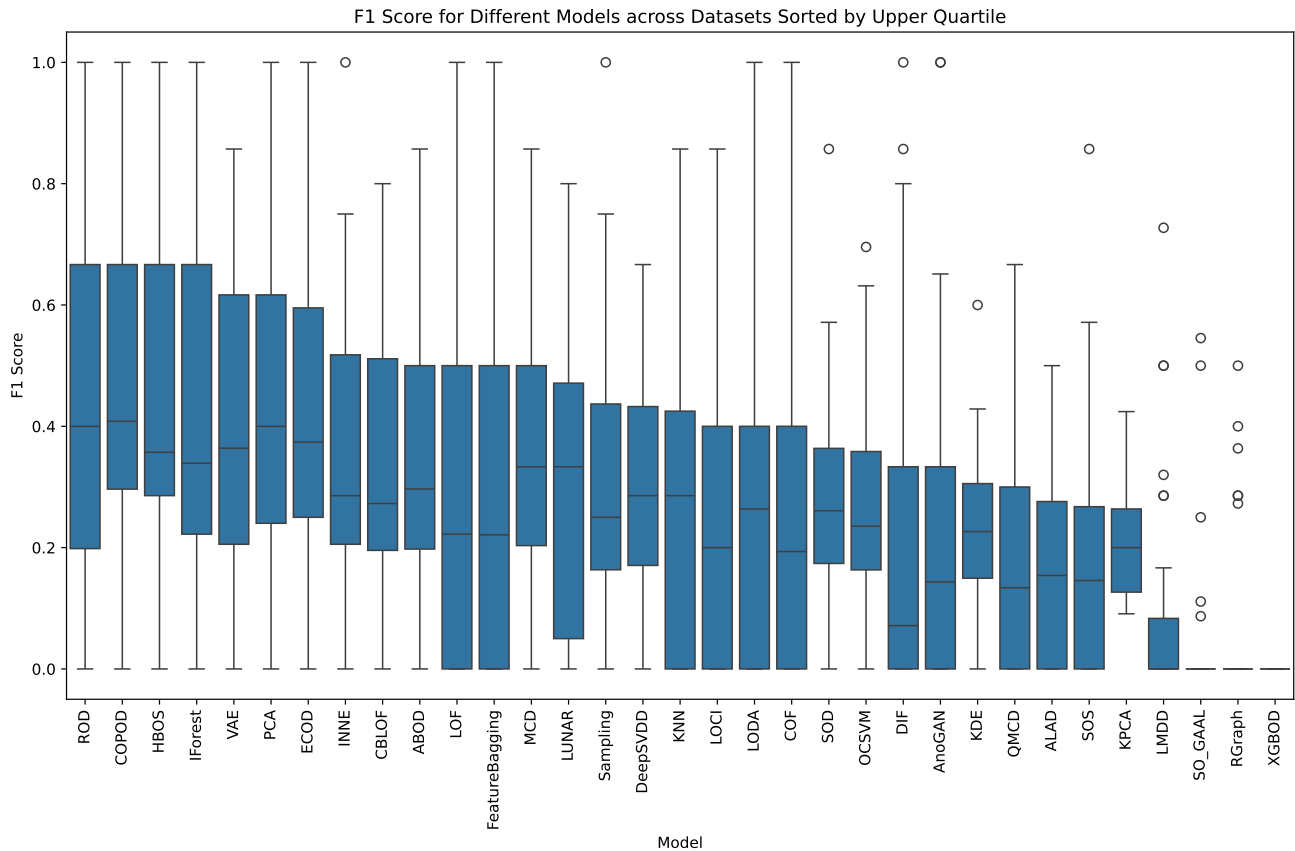


Figure 4. F1 Score for Different Models across Datasets Sorted by Upper Quartile

prehensively compare the best-performing models for each dataset based on the Average Precision (AP), ROC-AUC (AUC), and F1 Score. Table 4 presents the models that achieved the highest scores in these metrics across various healthcare datasets, underscoring the models' efficacy and suitability for specific data types. This table highlights the diversity in model performance across different datasets, illustrating that no single model uniformly excels across all performance metrics or datasets. For instance, models like COPOD and HBOS show top performance in multiple datasets for different metrics, suggesting their robustness and versatility in handling various data characteristics. Conversely, models like LODA and AnoGAN demonstrate exceptional performance in specific datasets, indicating their potential utility in tailored applications where their specific model architecture may align well with the dataset characteristics. The variability in the "best" model for each metric across the datasets also underscores the necessity of context-specific model selection in healthcare applications. For example, while AnoGAN excels in ROC-AUC Score for the appendicitis dataset, CD provides the best F1 Score, demonstrating the trade-offs that must be considered when prioritizing sensitivity, specificity, and precision.

Discussion

Our extensive evaluation of 37 outlier detection algorithms across 27 diverse healthcare datasets has provided valuable insights into the applicability, strengths, and limitations of these models in medical contexts. The datasets varied significantly in terms of dimensionality, sample size, artificially induced class imbalance, and feature types, which in turn influenced the performance of the outlier detection methods. This section discusses the key findings, highlighting how dataset characteristics impacted model efficacy and offers recommendations for practitioners.

Influence of Dataset Characteristics on Model Performance

The performance of outlier detection algorithms is intrinsically linked to the nature of the datasets they are applied to. High-dimensional datasets, such as the *Arrhythmia* dataset with 279 features and the *Darwin* dataset with 451 features, pose significant challenges due to the curse of dimensionality. In such cases, models that are specifically designed or adapted to handle high-dimensional data, like **ROD** and **HBOS**, demonstrated superior performance. For instance, **LODA**, which employs

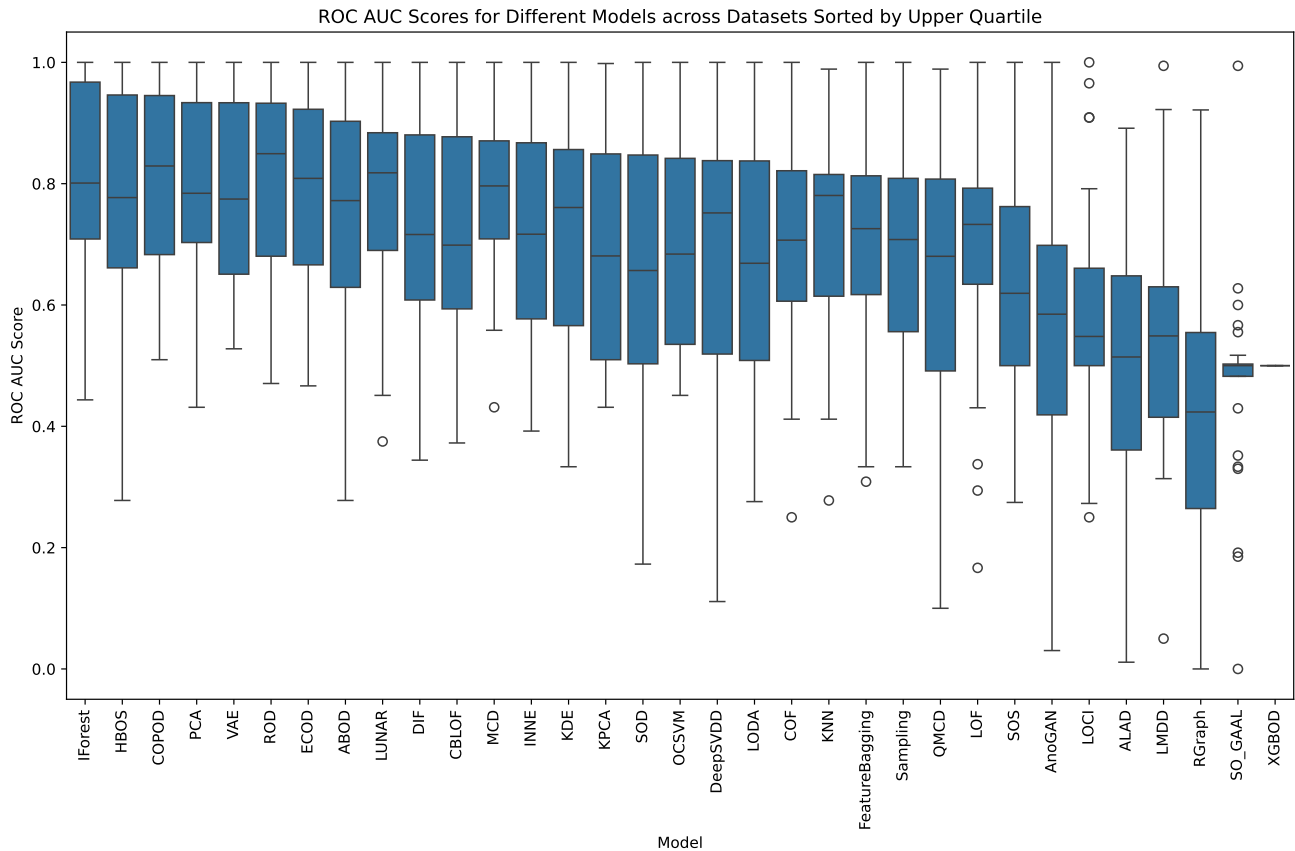


Figure 5. ROC-AUC Scores for Different Models across Datasets Sorted by Upper Quartile

random projections to reduce dimensionality, achieved perfect scores (AP, AUC, F1 Score all equal to 1.0) on the *Darwin* dataset (Table 4), indicating its effectiveness in managing high-dimensional spaces.

Conversely, datasets with low dimensionality, such as the *Haberman* dataset (4 features) and the *Appendicitis* dataset (8 features), showed different trends. Simpler models like **PCA** and **COF** performed reasonably well on these datasets. For example, **COF** achieved the best average precision and F1 Score on the *Haberman* dataset, suggesting that proximity-based methods are effective when the feature space is not overly complex.

Impact of Artificial Class Imbalance on Model Efficacy

Given the difficulty in obtaining rare disease data with naturally imbalanced classes, our methodology involved artificially reducing the number of positive cases to simulate class imbalance, reflecting the real-world scenario where certain medical conditions are rare. This intentional downsampling aimed to evaluate how well outlier detection models can identify anomalies (i.e., rare disease cases) within predominantly normal datasets.

The artificially induced class imbalance significantly impacted model performance. For instance, the *Lymphography* dataset was adjusted to have only 2 positive cases out of 148 samples (1.35% positive rate). Models like **ROD** and **ABOD**, which are robust to class imbalance, achieved perfect scores on this dataset. This underscores the importance of selecting models that can effectively handle datasets where anomalies are rare and deliberately underrepresented.

Probabilistic models like **COPOD** and **ECOD** also showed resilience to class imbalance, consistently performing well across datasets with low positive case ratios. Their ability to model the underlying data distribution without relying heavily on class labels makes them suitable for detecting outliers in imbalanced datasets.

Effectiveness of Different Algorithm Categories

Probabilistic models like **COPOD** and **ECOD** assume that outliers are observations that deviate significantly from the estimated probability distribution of the data. **COPOD**, in particular, showed strong performance across multiple datasets, including *Breast Cancer*, *Parkinson's Disease*, and *Thoracic Surgery*, often achieving the highest ROC-AUC scores. Its non-parametric

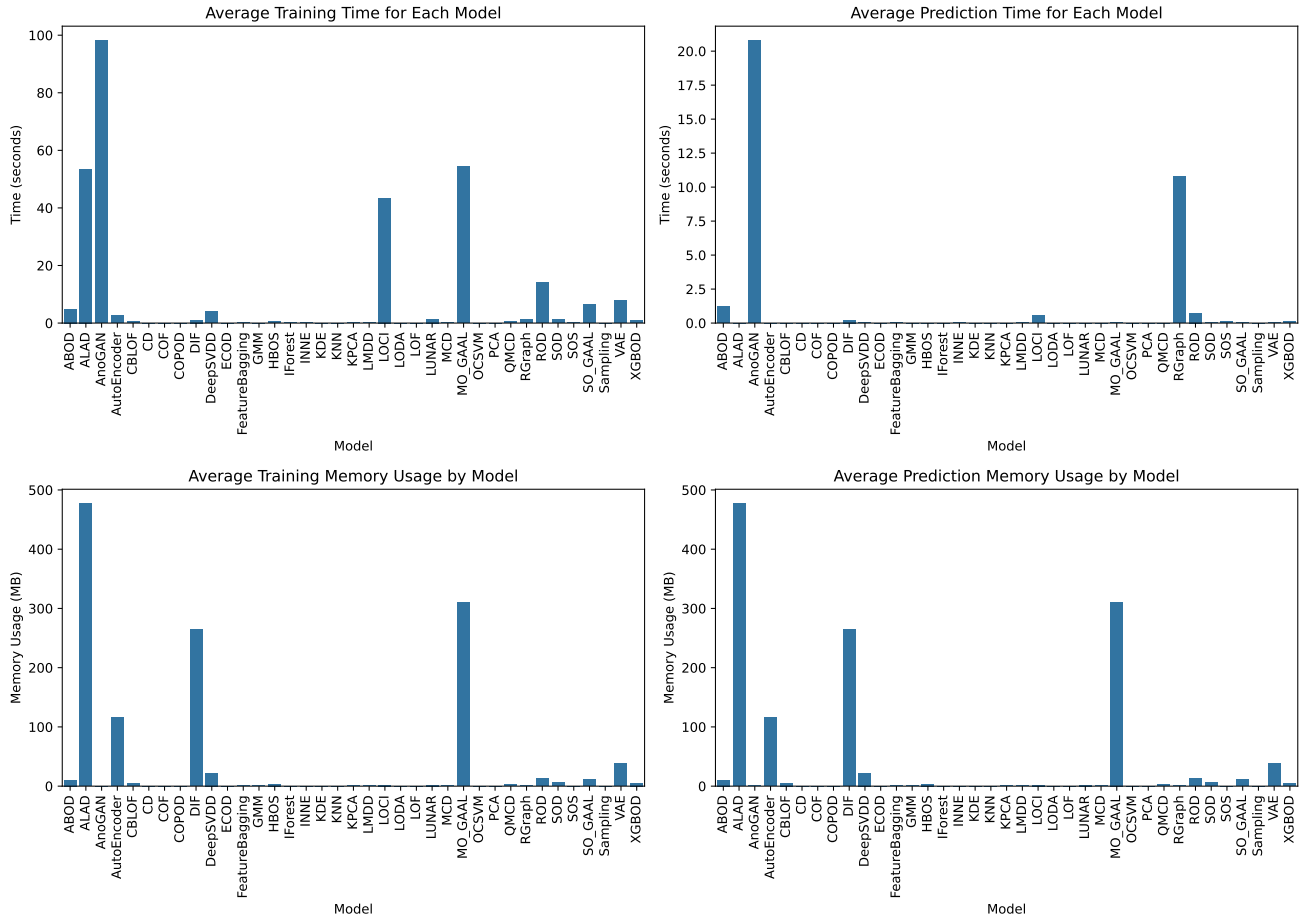


Figure 6. Average Training and Prediction Time and Memory Usage for Each Model

nature allows it to adapt to various data distributions without stringent assumptions, making it versatile for different healthcare datasets.

Linear models such as **PCA** and **MCD** performed well on datasets where linear relationships are prominent. For example, **PCA** achieved the best average precision on the *Spectfheart* dataset, which contains features derived from cardiac images. By capturing the principal components, **PCA** effectively identified deviations that signify outliers. However, these models may struggle with nonlinear patterns present in more complex datasets.

Proximity-based algorithms like **ROD** and **LOF** rely on the concept of distance or density in the feature space. **ROD** showed remarkable performance on several datasets, including *Heart Disease* and *Mammographic Mass*, indicating its strength in detecting outliers based on geometric properties. However, these methods can be sensitive to the choice of distance metrics and may not scale well with high-dimensional data unless combined with dimensionality reduction techniques.

Ensemble models like **IForest** and **LODA** combine multiple detection mechanisms to improve robustness. **IForest** was particularly effective on datasets like *Heart Failure* and *Hepatitis*, where it achieved high ROC-AUC scores. Its ability to isolate anomalies using random partitioning makes it suitable for datasets with complex, nonlinear relationships. However, ensemble methods often require more computational resources, which can be a limitation in resource-constrained environments.

Neural network models, including **AutoEncoder** and **DeepSVDD**, leverage deep learning architectures to model complex patterns. While these models showed potential, their performance was inconsistent across datasets. For instance, **AutoEncoder** achieved the best performance on the *CVD UAE* dataset, suggesting it can capture intricate relationships in cardiovascular data. However, on datasets like *Stroke*, neural network models did not outperform simpler methods, possibly due to overfitting on small sample sizes or insufficient training data.

Table 4. Performance Metrics of the Best Models for Each Dataset

Dataset	Best AP		Best AUC		Best F1	
	Model	AP	Model	AUC	Model	F1
Appendicitis	INNE	0.267	AnoGAN	0.765	CD	0.400
Breast cancer	HBOS	0.750	COPOD	1.000	HBOS	0.857
CVD UAE	AE ¹	0.321	AE ¹	0.972	AE ¹	0.487
Haberman	COF	0.208	PCA	0.704	COF	0.400
HCC data	COPOD	0.300	DeepSVDD	0.833	PCA	0.400
Heart	ROD	0.400	IForest	0.889	LODA	0.500
Mammographic	ROD	0.472	MCD	0.882	ROD	0.667
Pima	PCA	0.166	QMCD	0.750	PCA	0.353
Spectfheart	PCA	0.750	LODA	1.000	PCA	0.857
Stroke	HBOS	0.154	MCD	0.710	HBOS	0.333
Wisconsin	KNN	0.555	INNE	1.000	KNN	0.727
Heart failure	CD	0.115	IForest	0.758	ROD	0.222
Obesity	ABOD	0.594	QMCD	0.925	ABOD	0.727
Hepatitis	DIF	0.750	IForest	1.000	ECOD	0.857
Parkinsons	COPOD	0.750	FB ²	1.000	COPOD	0.857
Arrhythmia	SOS	0.289	MCD	0.846	SOS	0.500
Ilpd	LOF	0.346	LOF	0.752	LOF	0.500
Darwin	LODA	1.000	LODA	1.000	LODA	1.000
Lymphography	ROD	1.000	ABOD	1.000	ROD	1.000
Breast cancer coimbra	HBOS	1.000	MCD	1.000	HBOS	1.000
Cirrhosis	SOS	0.370	LODA	0.833	SOS	0.571
Vertebral	COPOD	0.488	CBLOF	0.917	MCD	0.667
Cervical cancer	ABOD	0.750	AE ¹	1.000	ABOD	0.857
Diabetic retinopathy	VAE	0.240	DeepSVDD	0.753	PCA	0.435
Bone marrow	Sampling	1.000	ABOD	1.000	Sampling	1.000
Thoracic surgery	SOD	0.185	COPOD	0.884	SOD	0.353
Thyroid	AnoGAN	1.000	ABOD	1.000	PCA	1.000

¹ AE: AutoEncoder² FB: FeatureBagging

Operational Considerations

Computational efficiency is critical for deploying models in clinical settings. Algorithms with high computational demands, such as **ABOD** and **LUNAR**, exhibited longer training and prediction times (Figure 6), making them less practical for real-time applications. In contrast, models like **HBOS** and **ECOD** offered a favorable balance between performance and efficiency, with low memory usage and fast execution times. For example, **HBOS** provided excellent results on the *Breast Cancer* and *Stroke* datasets while maintaining operational efficiency.

Trade-offs Between Sensitivity and Specificity

In healthcare, the costs associated with false positives and false negatives are significant. High sensitivity is crucial for early detection of diseases, while high specificity reduces unnecessary interventions. Our findings indicated that models like **PCA** and **MCD** achieved high sensitivity on datasets like *Spectfheart* and *Hepatitis*, but their specificity varied. **ROD** and **ECOD** maintained a better balance, achieving consistent F1 Scores across multiple datasets (Figure 4). This balance is essential in clinical scenarios where both missed diagnoses and false alarms carry substantial risks.

Case Studies of Specific Datasets

Arrhythmia Dataset

The *Arrhythmia* dataset, with its high dimensionality (279 features), represents a complex ECG signal analysis task. Models designed for high-dimensional data, such as **LODA** and **SOS**, performed better. **SOS** achieved the highest average precision and F1 Score on this dataset (Table 4). This suggests that models capable of handling sparse and high-dimensional data are more effective in detecting cardiac anomalies.

Diabetic Retinopathy Dataset

The *Diabetic Retinopathy* dataset includes features derived from image processing algorithms detecting microaneurysms and exudates. **VAE**, a neural network-based model, achieved the best average precision on this dataset, indicating its strength in modeling complex, high-dimensional data typical of medical imaging. However, **DeepSVDD** had the highest ROC-AUC score, highlighting that different models may capture different aspects of the data's underlying structure.

Obesity Dataset

The *Obesity* dataset contains both categorical and numerical features related to lifestyle and anthropometric measurements. **ABOD** performed exceptionally well, achieving the best average precision and F1 Score. This suggests that angle-based methods can effectively detect outliers in datasets where the relationships between features are non-linear and involve interactions between categorical and continuous variables.

Recommendations for Practitioners

Based on our findings, we offer the following recommendations:

- **Dataset Analysis:** Before selecting an outlier detection model, thoroughly analyze the dataset's characteristics, including dimensionality, sample size, feature types, and the degree of class imbalance, whether naturally occurring or artificially induced.
- **Model Selection:** Choose models that align with the dataset properties. For high-dimensional data, consider models like **LODA** or dimensionality reduction techniques combined with proximity-based methods. For datasets with significant class imbalance, probabilistic models like **COPOD** or isolation-based models like **IForest** are recommended.
- **Operational Constraints:** In real-time or resource-limited settings, prioritize models with lower computational demands, such as **HBOS** and **ECOD**, which provide a good balance between accuracy and efficiency.
- **Interpretability:** Favor models that offer interpretability, especially in clinical contexts where understanding the reasoning behind an outlier detection is crucial for decision-making. Linear models and simpler probabilistic methods are generally more interpretable.
- **Validation and Testing:** Use cross-validation and consider multiple performance metrics, including ROC-AUC, average precision, F1 Score, sensitivity, and specificity, to comprehensively evaluate model performance.
- **Adaptability:** Be prepared to combine models or use ensemble approaches when a single model does not meet all the performance requirements. Hybrid models can sometimes capture complementary aspects of the data.

Conclusion

This study comprehensively evaluated the effectiveness of 37 outlier detection algorithms across 27 healthcare datasets, providing key insights into their performance and suitability for specific healthcare applications. The findings highlight the importance of choosing the right outlier detection method based on dataset characteristics, desired performance metrics, and computational limitations. We evaluated each model using performance metrics such as ROC AUC, average precision, F1 Score, sensitivity, and specificity, as well as operational metrics like training time, prediction time, and memory usage, which are critical for practical use.

Our results show significant variability in model performance across different datasets, emphasizing the importance of context when selecting outlier detection methods. Models like **COPOD**, **HBOS**, and **IForest** consistently performed well across multiple metrics, making them suitable for a range of healthcare applications. However, some models, like **LODA** and **AnoGAN**, excelled in specific datasets but did not generalize well across others, suggesting the need for specialized approaches in certain scenarios.

Additionally, the operational metrics offered valuable insights into the feasibility of deploying these models in real-world healthcare settings, where computational resources and quick response times are often constrained. Models that strike a balance between high accuracy and reasonable computational demands are ideal for real-time applications, ensuring they detect anomalies effectively while operating within healthcare system limits.

These findings underline the need for healthcare professionals and data scientists to carefully consider both the performance and operational characteristics of outlier detection models. Choosing the appropriate model based on both accuracy and efficiency can enhance healthcare systems' ability to detect and respond to anomalies, improving patient outcomes and overall operational efficiency. Furthermore, this study contributes to the ongoing discussion on the role of outlier detection methods in healthcare, a field where accurate outlier detection can have a significant impact on patient care. By comparing a variety of models, this research helps practitioners and researchers make informed decisions about which models to use based on specific requirements and constraints.

Limitations and Future Work

This study has several limitations that affect the evaluation of outlier detection models. First, the datasets used are relatively small, with a maximum of 2111 samples. This size constraint limits the effectiveness of deep learning models like DeepSVDD, often resulting in suboptimal performance. Second, some models encountered training errors, which prevented a full evaluation of their performance alongside other models. As a result, we excluded these models from the analysis, which may have introduced bias in the overall comparisons.

Despite these limitations, this study offers valuable insights into the performance of various outlier detection models across healthcare datasets. However, the fast-changing nature of healthcare data and outlier detection methods presents many opportunities for future research. Future studies should use a broader range of datasets, including larger and more complex ones, to enhance the robustness and applicability of outlier detection models. Expanding the dataset variety would also help assess how well these models scale and adapt to different healthcare environments and patient populations.

Acknowledgments

This research is partially supported by ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10.

Author contributions

J.R., Z.R., and N.Z. conceived the experiment(s). J.R. conducted the experiment(s). J.R. and N.B. analysed the results. All authors reviewed the manuscript.

Competing interests

The author(s) declare no competing interests.

Data availability

The datasets analysed during the current study are available in our GitHub repository with links to the original repository where we extracted each dataset from, <https://github.com/jaloliddin-rus/outlier-detection-healthcare.git>.

References

1. Goldstein, M. & Uchida, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE* **11**, 1–31, DOI: [10.1371/journal.pone.0152173](https://doi.org/10.1371/journal.pone.0152173) (2016).
2. Campos, G. O. *et al.* *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study*, vol. 30 (Springer US, 2016).
3. Aggarwal, C. C. & Yu, P. S. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, SIGMOD '01, 37–46, DOI: [10.1145/375663.375668](https://doi.org/10.1145/375663.375668) (Association for Computing Machinery, New York, NY, USA, 2001).
4. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowl. Data Eng.* **21**, 1263–1284, DOI: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239) (2009).
5. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Medicine* **23**, 89–109, DOI: [https://doi.org/10.1016/S0933-3657\(01\)00077-X](https://doi.org/10.1016/S0933-3657(01)00077-X) (2001).
6. Xu, X., Liu, H. & Yao, M. Recent Progress of Anomaly Detection. *Complexity* **2019**, DOI: [10.1155/2019/2686378](https://doi.org/10.1155/2019/2686378) (2019).
7. Gaspar, J., Catumbela, E., Marques, B. & Freitas, A. A systematic review of outliers detection techniques in medical data: Preliminary study. *HEALTHINF 2011 - Proc. Int. Conf. on Heal. Informatics* 575–582, DOI: [10.5220/0003168705750582](https://doi.org/10.5220/0003168705750582) (2011).
8. Emmott, A., Das, S., Dietterich, T., Fern, A. & Wong, W.-K. A meta-analysis of the anomaly detection problem (2016). [1503.01158](https://arxiv.org/abs/1503.01158).
9. Domingues, R., Filippone, M., Michiardi, P. & Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit.* **74**, 406–421, DOI: [10.1016/j.patcog.2017.09.037](https://doi.org/10.1016/j.patcog.2017.09.037) (2018).
10. Steinbuss, G. & Böhm, K. Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data. *ACM Transactions on Knowl. Discov. from Data* **15**, DOI: [10.1145/3441453](https://doi.org/10.1145/3441453) (2021).

11. Soenen, J. *et al.* The Effect of Hyperparameter Tuning on the Comparative Evaluation of Unsupervised Anomaly Detection Methods. *Proc. KDD'21 Work. on Outlier Detect. Descr.* 1–9 (2021).
12. Han, S., Hu, X., Huang, H., Jiang, M. & Zhao, Y. ADBench: Anomaly Detection Benchmark. *Adv. Neural Inf. Process. Syst.* **35**, DOI: [10.2139/ssrn.4266498](https://doi.org/10.2139/ssrn.4266498) (2022).
13. UCI Machine Learning Repository.
14. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J. & García, S. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *J. Multiple Valued Log. Soft Comput.* (2011).
15. Eteng, I. *et al.* UCTH Breast Cancer Dataset, DOI: [10.17632/63FPBC9CM4.2](https://doi.org/10.17632/63FPBC9CM4.2) (2023).
16. Al-Shamsi, S., Regmi, D. & Govender, R. D. Incidence of cardiovascular disease and its associated risk factors in at-risk men and women in the United Arab Emirates: a 9-year retrospective cohort study. *BMC cardiovascular disorders* **19**, 148, DOI: [10.1186/s12872-019-1131-2](https://doi.org/10.1186/s12872-019-1131-2) (2019).
17. Smith, J. W., Everhart, J., Dickson, W., Knowler, W. & Johannes, R. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proc. Annu. Symp. on Comput. Appl. Med. Care* 261 (1988).
18. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
19. Emmott, A. F., Das, S., Dietterich, T., Fern, A. & Wong, W. K. Systematic construction of anomaly detection benchmarks from real data. *Proc. ACM SIGKDD Work. on Outlier Detect. Descr. ODD 2013* 16–21, DOI: [10.1145/2500853.2500858](https://doi.org/10.1145/2500853.2500858) (2013).
20. Janssens, J. *Outlier selection and one-class classification*. Ph.D. thesis, Tilburg University (2013). Series: TiCC Ph.D. Series Volume: 27.
21. Fang, K. T. & Ma, C. X. Wrap-Around L2-Discrepancy of Random Sampling, Latin Hypercube and Uniform Designs. *J. Complex.* **17**, 608–624, DOI: [10.1006/JCOM.2001.0589](https://doi.org/10.1006/JCOM.2001.0589) (2001).
22. Pavlidou, M. & Zioutas, G. Kernel density outlier detector. In *Springer Proceedings in Mathematics and Statistics*, vol. 74, 241–250 (Springer New York LLC, 2014).
23. Aggarwal, C. C. An Introduction to Outlier Analysis. In *Outlier Analysis*, 1–34 (Springer International Publishing, Cham, 2017).
24. Kriegel, H.-P., Schubert, M. & Zimek, A. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 444–452, DOI: [10.1145/1401890.1401946](https://doi.org/10.1145/1401890.1401946) (ACM, New York, NY, USA, 2008).
25. Li, Z., Zhao, Y., Botta, N., Ionescu, C. & Hu, X. COPOD: Copula-Based Outlier Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, vol. 2020-Novem, 1118–1123, DOI: [10.1109/ICDM50108.2020.00135](https://doi.org/10.1109/ICDM50108.2020.00135) (IEEE, 2020).
26. Li, Z. *et al.* ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowl. Data Eng.* **35**, 12181–12193, DOI: [10.1109/TKDE.2022.3159580](https://doi.org/10.1109/TKDE.2022.3159580) (2023).
27. Sugiyama, M. & Borgwardt, K. M. Rapid Distance-Based Outlier Detection via Sampling. *Adv. Neural Inf. Process. Syst.* **26** (2013).
28. Cook, R. D. Detection of Influential Observation in Linear Regression. *Technometrics* **19**, 15, DOI: [10.2307/1268249](https://doi.org/10.2307/1268249) (1977).
29. Arning, A., Agrawal, R. & Raghavan, P. A linear method for deviation detection in large databases. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, 164–169 (AAAI Press, 1996).
30. Rousseeuw, P. J. & Driessen, K. V. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* **41**, 212–223, DOI: [10.1080/00401706.1999.10485670](https://doi.org/10.1080/00401706.1999.10485670) (1999).
31. Shyu, M.-L., Chen, S., Sarinnapakorn, K. & Chang, L. A novel anomaly detection scheme based on principal component classifier (2003).
32. Hoffmann, H. Kernel PCA for novelty detection. *Pattern Recognit.* **40**, 863–874, DOI: [10.1016/j.patcog.2006.07.009](https://doi.org/10.1016/j.patcog.2006.07.009) (2007).
33. Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J. & Platt, J. Support vector method for novelty detection. In Solla, S., Leen, T. & Müller, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 12 (MIT Press, 1999).
34. Almarideny, Y., Boujnah, N. & Cleary, F. A Novel Outlier Detection Method for Multivariate Data. *IEEE Transactions on Knowl. Data Eng.* **34**, 4052–4062, DOI: [10.1109/TKDE.2020.3036524](https://doi.org/10.1109/TKDE.2020.3036524) (2022).

35. Ramaswamy, S., Rastogi, R. & Shim, K. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Rec.* **29**, 427–438, DOI: [10.1145/335191.335437](https://doi.org/10.1145/335191.335437) (2000).
36. He, Z., Xu, X. & Deng, S. Discovering cluster-based local outliers. *Pattern Recognit. Lett.* **24**, 1641–1650, DOI: [10.1016/S0167-8655\(03\)00003-5](https://doi.org/10.1016/S0167-8655(03)00003-5) (2003).
37. Goldstein, M., track, A. D. K.-. p., demo & 2012, u. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *CiteseerM Goldstein, A DengelKI-2012: poster demo track, 2012*.
38. Breunig, M. M., Kriegel, H. P., Ng, R. T. & Sander, J. LOF: Identifying Density-Based Local Outliers. In *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104, DOI: [10.1145/342009.335388](https://doi.org/10.1145/342009.335388) (ACM, New York, NY, USA, 2000).
39. Papadimitriou, S., Kitagawa, H., Gibbons, P. B. & Faloutsos, C. LOCI: Fast outlier detection using the local correlation integral. In *Proceedings - International Conference on Data Engineering*, 315–326, DOI: [10.1109/ICDE.2003.1260802](https://doi.org/10.1109/ICDE.2003.1260802) (IEEE, 2003).
40. Tang, J., Chen, Z., Fu, A. W. C. & Cheung, D. W. Enhancing effectiveness of Outlier detections for low Density Patterns. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2336, 535–548 (Springer, Berlin, Heidelberg, 2002).
41. Kriege, H. P., Kroger, P., Schubert, E. & Zimek, A. Outlier detection in axis-parallel subspaces of high dimensional data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5476 LNAI, 831–838 (Springer, Berlin, Heidelberg, 2009).
42. Xu, H., Pang, G., Wang, Y. & Wang, Y. Deep Isolation Forest for Anomaly Detection. *IEEE Transactions on Knowl. Data Eng.* **35**, 12591–12604, DOI: [10.1109/TKDE.2023.3270293](https://doi.org/10.1109/TKDE.2023.3270293) (2023).
43. Bandaragoda, T. R. *et al.* Isolation-based anomaly detection using nearest-neighbor ensembles. *Comput. Intell.* **34**, 968–998, DOI: [10.1111/coin.12156](https://doi.org/10.1111/coin.12156) (2018).
44. Lazarevic, A. & Kumar, V. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 157–166, DOI: [10.1145/1081870.1081891](https://doi.org/10.1145/1081870.1081891) (ACM, New York, NY, USA, 2005).
45. Pevný, T. Loda: Lightweight on-line detector of anomalies. *Mach. Learn.* **102**, 275–304, DOI: [10.1007/S10994-015-5521-0/FIGURES/10](https://doi.org/10.1007/S10994-015-5521-0/FIGURES/10) (2016).
46. Zhao, Y. & Hryniewicki, M. K. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, vol. 2018-July, 1–8, DOI: [10.1109/IJCNN.2018.8489605](https://doi.org/10.1109/IJCNN.2018.8489605) (IEEE, 2018).
47. Ruff, L. *et al.* Deep One-Class Classification (2018).
48. Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2022). [1312.6114](https://arxiv.org/abs/1312.6114).
49. Zenati, H., Romain, M., Foo, C.-S., Lecouat, B. & Chandrasekhar, V. Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, 727–736, DOI: [10.1109/ICDM.2018.00088](https://doi.org/10.1109/ICDM.2018.00088) (2018).
50. Liu, Y. *et al.* Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Transactions on Knowl. Data Eng.* **32**, 1–1, DOI: [10.1109/TKDE.2019.2905606](https://doi.org/10.1109/TKDE.2019.2905606) (2019).
51. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, 146–147 (Springer Verlag, 2017).
52. Goodge, A., Hooi, B., Ng, S.-K. & Ng, W. S. LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks. *Proc. AAAI Conf. on Artif. Intell.* **36**, 6737–6745, DOI: [10.1609/aaai.v36i6.20629](https://doi.org/10.1609/aaai.v36i6.20629) (2022).
53. You, C., Robinson, D. P. & Vidal, R. Provable self-representation based outlier detection in a union of subspaces (2017). [1704.03925](https://arxiv.org/abs/1704.03925).
54. Zhao, Y., Nasrullah, Z. & Li, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *J. Mach. Learn. Res.* **20**, 1–7 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFilesforBenchmarkingOutlierDetectionAlgorithmsinHealthcareAnAnalysisAcrossDiverseMedicalDatasets.pdf](#)