10.48047/jocaaa.2024.33.02.27

A Comparative Study of Outlier Detection Methods in Heart Disease Data

K. Senthamarai Kannan[1] and D. Kabinath[2]
[1,2]Department of Statistics,
Manonmaniam Sundaranar University, Tirunelveli – 627 012.
Email: senkannan2002@gmail.com, kabi130198@gmail.com

Abstract

Outlier is defined as an observation that deviates significantly from other observations. Identifying outliers can lead to the discovery of valuable and meaningful information. Over the last few decades, there has been extensive research into outlier detection. The most advanced data mining techniques partially address this issue, but not entirely, and can be enhanced by taking a more direct approach to the issue. Outlier detection can result in the discovery of unexpected facts in fields such as credit card fraud detection, calling card fraud detection, criminal behavior detection, network intrusion detection, and so on. In this study, we applied and compared distance-based and density-based outlier detection techniques on heart disease datasets, and evaluated their impact on classification performance using Naive Bayes and Support Vector Machine algorithms.

Key words: Outlier detection, Heart disease, k-nearest neighbors method, Local outlier factor method.

Introduction

Data mining is the process of obtaining valid, not previously known, and finally interpretable information from extremely large datasets to support organizational decision-making. There are several problems that arise while obtaining data from large datasets, such as data redundancy, invalid attribute values, missing records, and the presence of outliers. The data industry has a very huge amount of data. The data will be valuable only if it is transformed into useful information. Processing this immense data is required in order to obtain useful insights from it.

Outlier is defined as an observation that deviates so significantly from other observations that it raises suspicions that it was caused by a different mechanism than other interpretations. Outlier detection can lead to the discovery of useful knowledge and has a variety of practical applications, including public safety, transportation, public health, and location-based services. This paper focuses on

10.48047/jocaaa.2024.33.02.27

outlier detection approaches from a data mining perspective. The underlying idea is to analyze and compare the achieving mechanisms of those approaches in order to determine which approach is superior based on noisy data.

Related Works

Detection of outliers is at the core of the medical data preprocessing, particularly in disease diagnosis applications, where outlier or noisy data will drastically change classification performance. Two of the most common methodologies employed are distance-based and density-based.

Distance-based methods, represented by the k-Nearest Neighbors (k-NN) algorithm, define outliers as those that are far from their neighboring data points. A point is considered an outlier if it is far from its k nearest neighbors. k-NN outlier detection was used by Mahalakshmi and Govindarajan [5] on the PIMA Indian diabetes dataset and established that classifier accuracy increased after the removal of the best-scoring outliers before training Naive Bayes and Support Vector Machine (SVM) classifiers.

Density-based techniques, like the Local Outlier Factor (LOF), detect outliers by comparing the local density difference of an individual point to those of its surrounding points. The technique is particularly useful on high-dimensional and complex data distributions with heterogeneous regions of density, allowing the detection of local outliers but not global outliers. Breunig et al. [1] originally described the LOF approach, demonstrating its effectiveness on high-dimensional and complex data distributions.

To improve detection efficiency and computation, hybrid approaches that integrate clustering with outlier detection techniques have been introduced. Surekha and Dongre [3] introduced an algorithm named I-CLARANS, which integrates partition-based clustering algorithms such as PAM and CLARA with distance-based outlier detection techniques. This method significantly improved computation time with higher accuracy in the detection of abnormal records.

While these methods have been extensively applied to datasets for diabetes, credit card fraud, and network intrusion, few studies have focused on heart disease-related datasets. Since cardiovascular diagnosis at an early stage becomes clinically relevant, one must understand the effect of outlier detection on model performance in this specific scenario. Pathan et al. [7] and Idri et al. [6] mention that preprocessing methods like outlier removal and feature selection are central to constructing robust prediction models for heart disease.

10.48047/jocaaa.2024.33.02.27

In this research, we investigate the impacts of distance-based (k-NN) and density-based (LOF) outlier detection algorithms on heart disease datasets. We investigate their impacts on classification accuracy when utilized with SVM and Naive Bayes classifiers and conduct a performance comparison of each algorithm in medical diagnostic contexts.

Datasets Description

The dataset used in this study is the Heart Disease dataset, which has been extensively studied and is considered challenging. It contains 120 instances (44.4%) classified as class '1' and 150 instances (55.6%) classified as class '0'. Table 1 provides a description of the dataset used.

Table 1 Description of dataset

Property	Value
Number of Samples	270
Number of Attributes	13
Number of Classes	2
Type of Attributes	Numeric
Type of Class Attribute	Binomial (presence/absence of heart disease)

Heart Noise Datasets

10.48047/jocaaa.2024.33.02.27

The standard classification task involves making generalizations from a set of training examples. To evaluate the robustness of outlier detection techniques, artificial noise was introduced into the heart disease dataset. The types and levels of noise added are summarized in Table 2.

Table 2 Types of Heart Noise Dataset

Dataset Name	% of Noise Added	Description
Heart Noise Dataset 0 (HND0)	0%	Original dataset
Heart Noise Dataset 1 (HND1)	5%	5% attribute noise added
Heart Noise Dataset 2 (HND2)	10%	10% attribute noise added
Heart Noise Dataset 3 (HND3)	15%	15% attribute noise added
Heart Noise Dataset 4 (HND4)	20%	20% attribute noise added

The dataset has been contaminated with attribute noise. As detailed in Table 2, five distinct types of datasets are utilized: the original dataset and datasets featuring 5%, 10%, 15%, and 20% attribute noise.

Outlier Detection Methods

Identifying outliers is a crucial preprocessing step for extracting valuable insights from data. By ensuring datasets are devoid of outliers, we can significantly minimize ambiguity and uncertainty in the analysis. Various experiments were conducted on multiple heart disease datasets, as detailed in Table 2. These experiments operated under the assumption that the datasets contained labels indicating normal behavior or class affiliation. In this research, two main techniques for outlier detection were employed: one based on distance and the other on density. Both approaches assess each instance within the dataset and assign a score that indicates the likelihood of it being an outlier, facilitating the comparison and evaluation of their effectiveness.

Distance based Outlier Detection

There are various distance-based techniques, and the k-NN technique for outlier detection is utilized in this study. It is a technique of outlier detection based

10.48047/jocaaa.2024.33.02.27

on how close an object is to its k nearest neighbors, and the object to be detected is one of the neighbors. For every item in the data set, the items that are its neighbors within a certain distance are counted. Those items with the least number of neighbors within this certain radius are the noise candidates. The data set is then sorted in ascending order by the number of neighbors for every item. The first n items, having the least number of neighbors, are marked as noise and then removed from the data set. The main parameters that define this k-NN distance-based technique are the distance function and the parameter "k." In the implementation here, the Euclidean distance function is used.

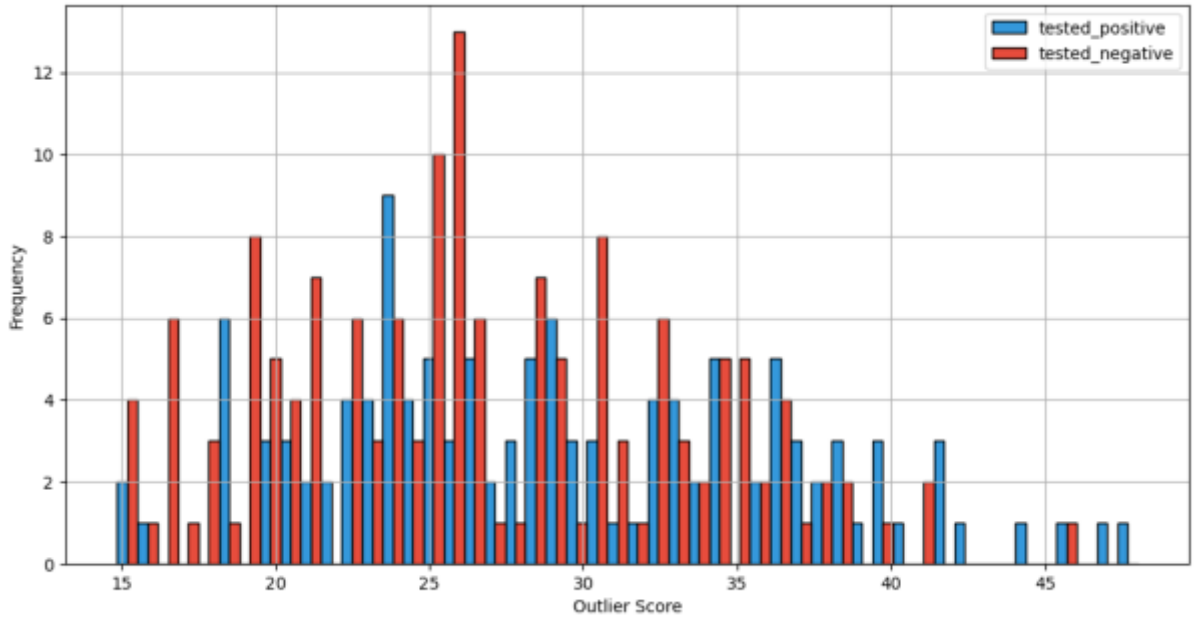


Figure 1: Distance based outlier score distribution of HND1

For HND1, the distance based outlier detection technique generates outlier scores ranging between 15 and 48. The resultant histogram for HND1 is shown in Figure 1. Among the outlier scores calculated, it can be seen that very few instances have highest outlier score of 45.

10.48047/jocaaa.2024.33.02.27

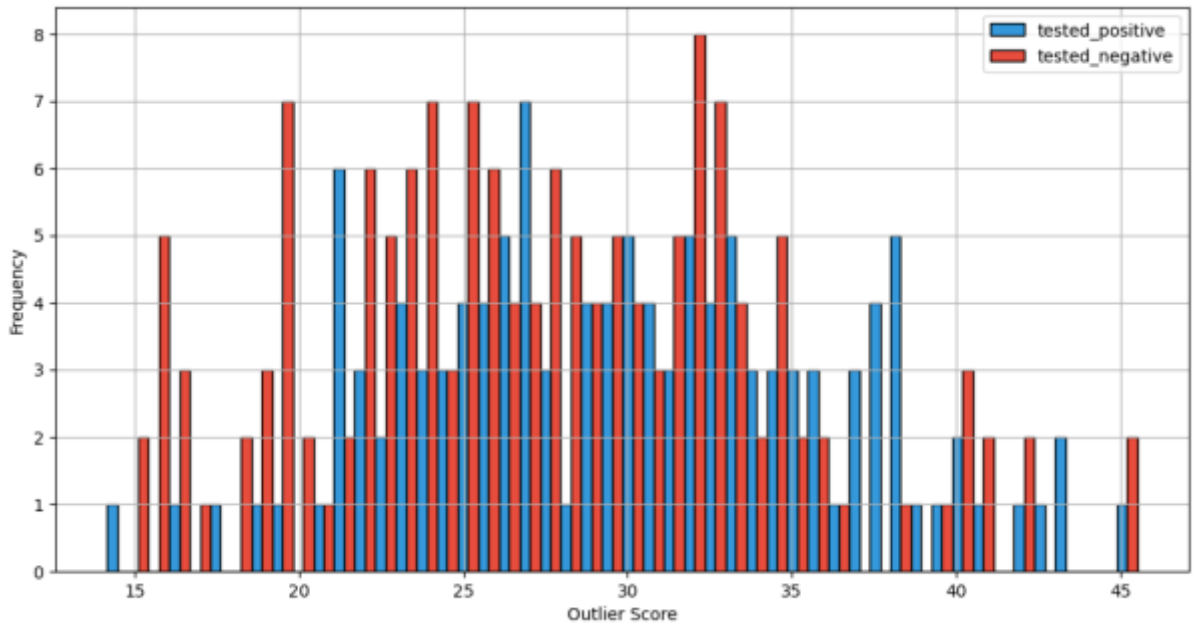


Figure 2: Distance based outlier score distribution of HND2

For HND2, the distance based outlier detection technique generates outlier scores ranging between 14 and 45. The resultant histogram for HND1 is shown in Figure 2. In Figure 2, for HND2, the maximum number of instances are distributed evenly for both positive and negative class instances. The peak occurs at outlier score value of 32.

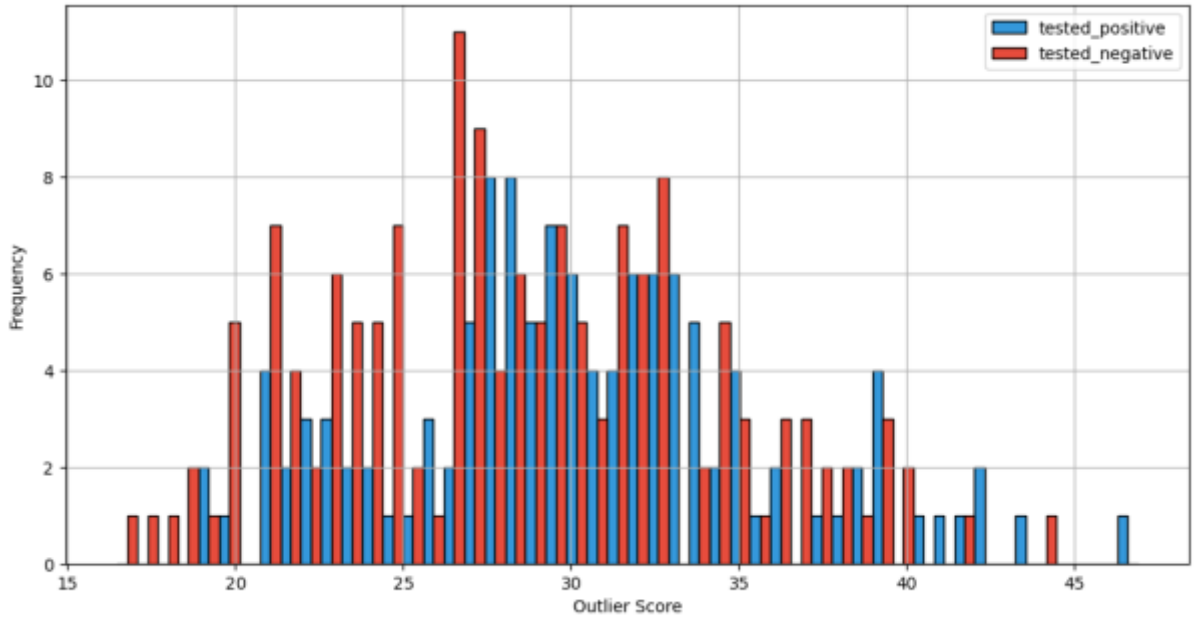


Figure 3: Distance based outlier score distribution of HND3

10.48047/jocaaa.2024.33.02.27

For HND3, the distance based outlier detection technique generates outlier scores ranging between 17 and 47. The resultant histogram for HND3 is shown in Figure 3. Among the outlier scores calculated, it can be seen that very few instances have highest outlier score of 43.

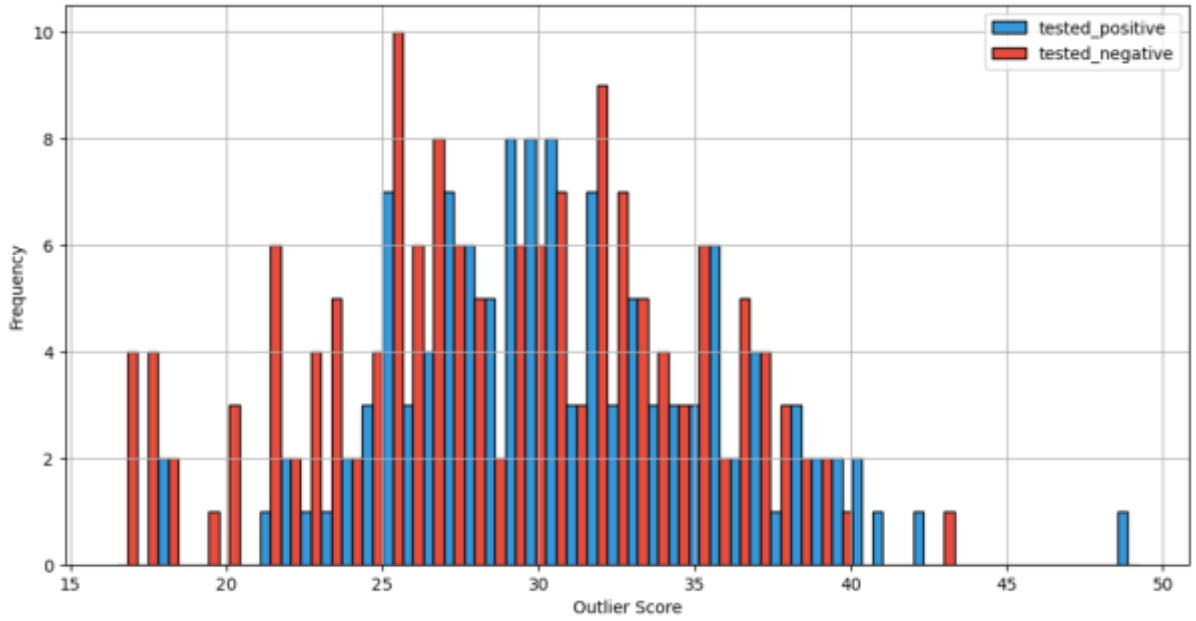


Figure 4: Distance based outlier score distribution of HND4

From Figure 4, it can be inferred that the outlier values are distributed between 17 and 49. In Figure 4, for HND4, the maximum number of instances are distributed with outlier score of around 25-35 for both positive and negative class instances. Among the outlier scores calculated, Among the outlier scores calculated, it can be seen that very few instances have highest outlier score of 40.

Density based Outlier Detection

Density-based outlier detection makes use of the density distribution of data points within the dataset. This technique measures outliers using a (LOF). The LOF is the ratio of an object's local density to that of its nearest neighbor. Outliers are data points with a higher LOF. The algorithm for computing the density-based local outlier factor in a database is based on minPTs. The original LOF parameter was called "minPts", but for consistency within ELKI, it is now called "k". The "k" value indicates the number of nearest neighbors whose distance is used to estimate density. In this approach, the Euclidean distance function is used. For the datasets, HND1, HND2, HND3, and HND4 used, the calculated outlier values are sketched in Figure 5 – Figure 8 in the form of a histogram.

10.48047/jocaaa.2024.33.02.27

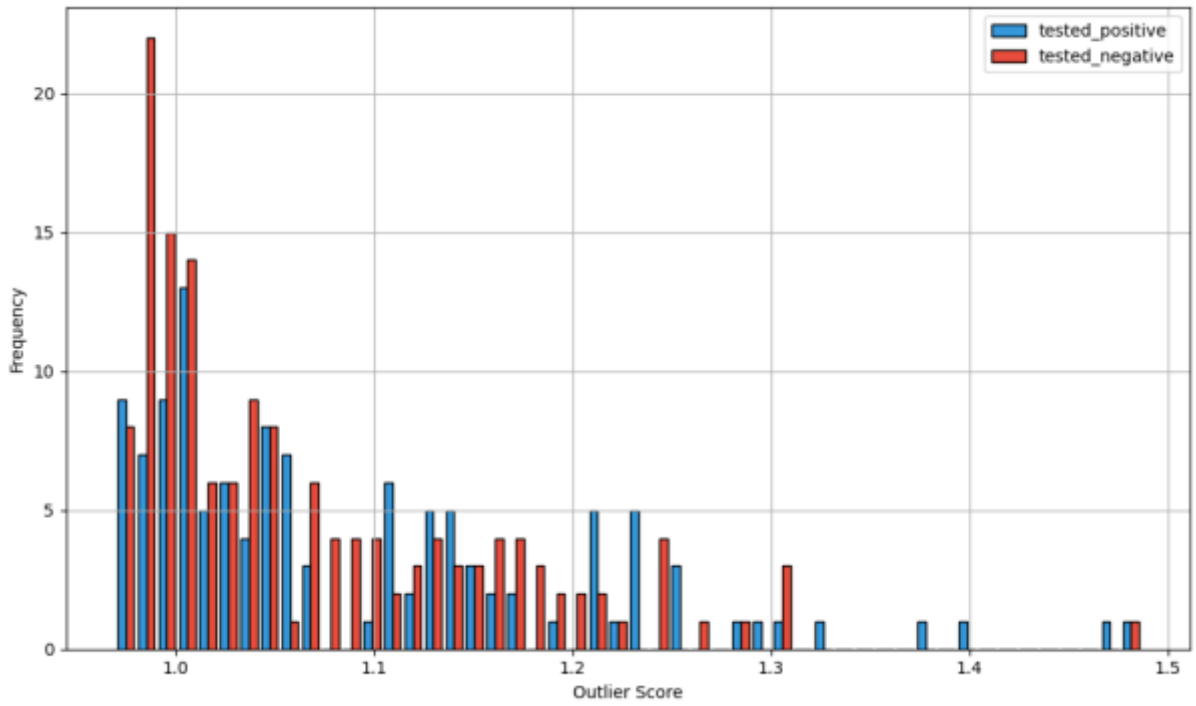
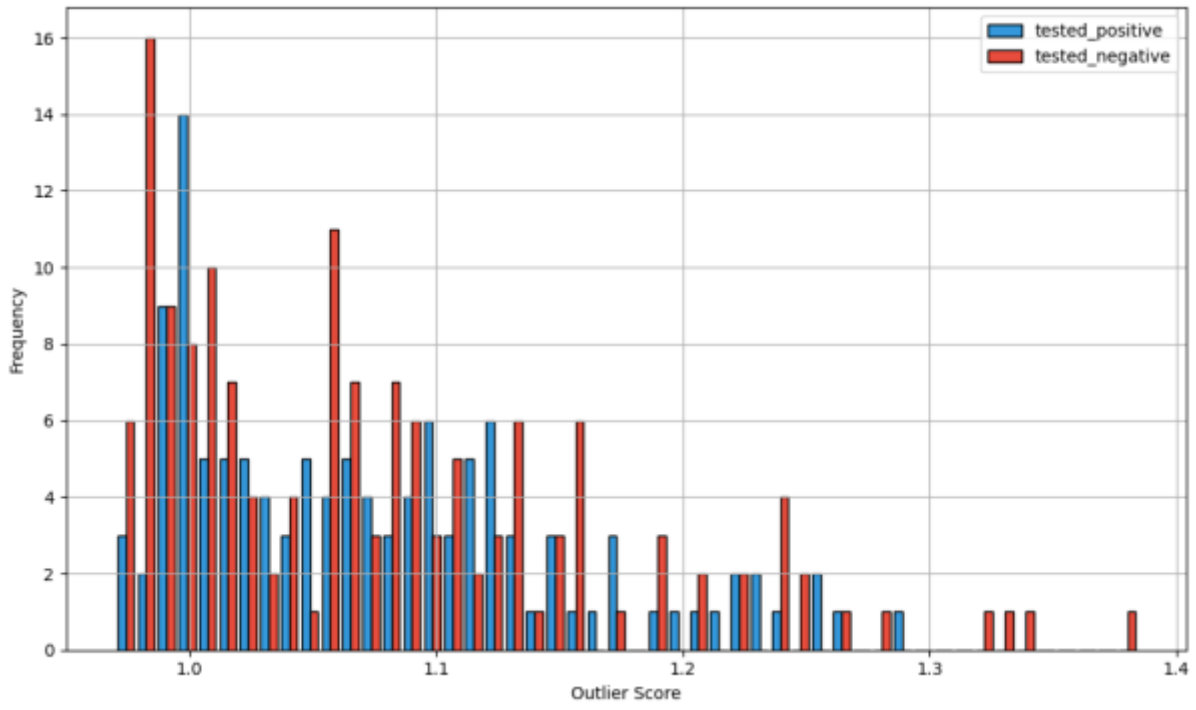


Figure 5: Density based outlier score distribution of HND1

Figure 5 presents a histogram showing the distribution of LOF outlier scores for the HND1 dataset using a density-based outlier detection approach. The outlier scores range from approximately 0.9 to 1.48, with the highest frequency occurring near a score of 0.95, indicating that most data points are considered normal. Only a small number of instances exhibit outlier scores above 1.3, suggesting the presence of a few strong outliers in the dataset.



10.48047/jocaaa.2024.33.02.27

Figure 6: Density based outlier score distribution of HND2

From Figure 6, it can be inferred that the outlier values are distributed between 0.9 and 1.4. In Figure 6, for HND2, the maximum number of instances are distributed around outlier score of around 1-1.3 for both positive and negative class instances. Among the outlier scores calculated, the top 10 outliers are to be identified from outlier score greater than 1.3.

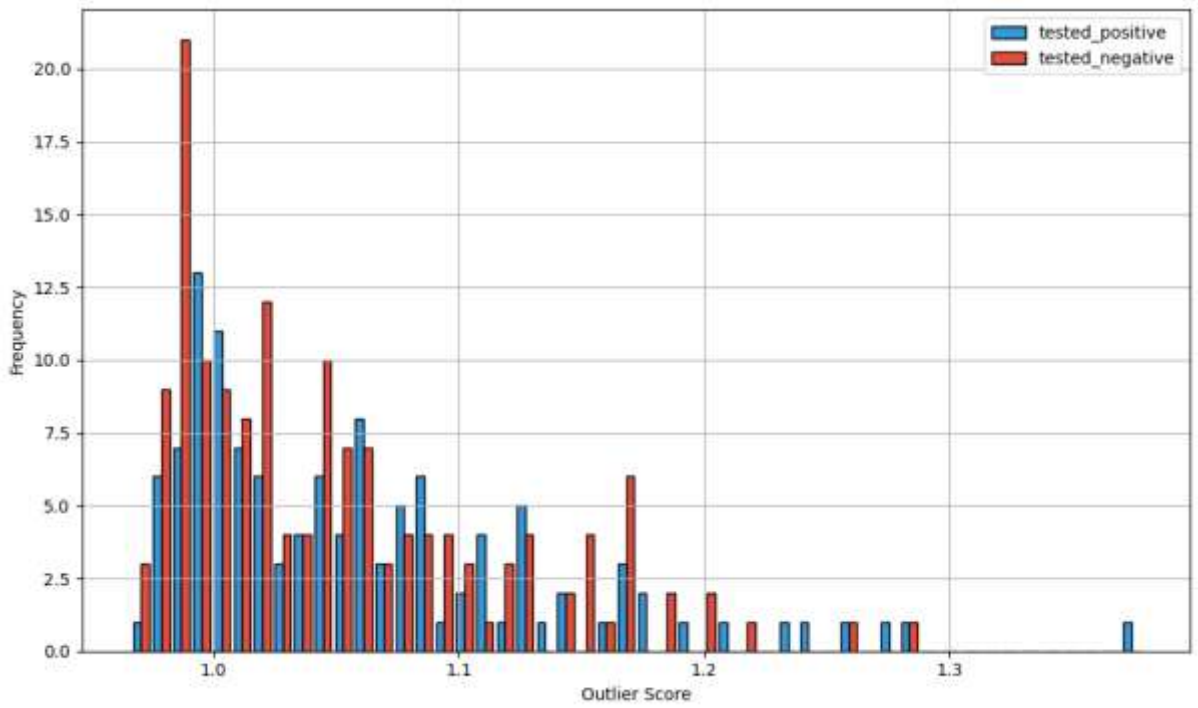


Figure 7: Density based outlier score distribution of HND3

From Figure 7, it can be inferred that the outlier values are distributed between 0.97 and 1.37. In Figure 6, for HND3, the maximum number of instances are distributed around outlier score of around 0.99 – 1.2 for both positive and negative class instances. Among the outlier scores calculated, the top 10 outliers are to be identified from outlier score greater than 1.2.

10.48047/jocaaa.2024.33.02.27

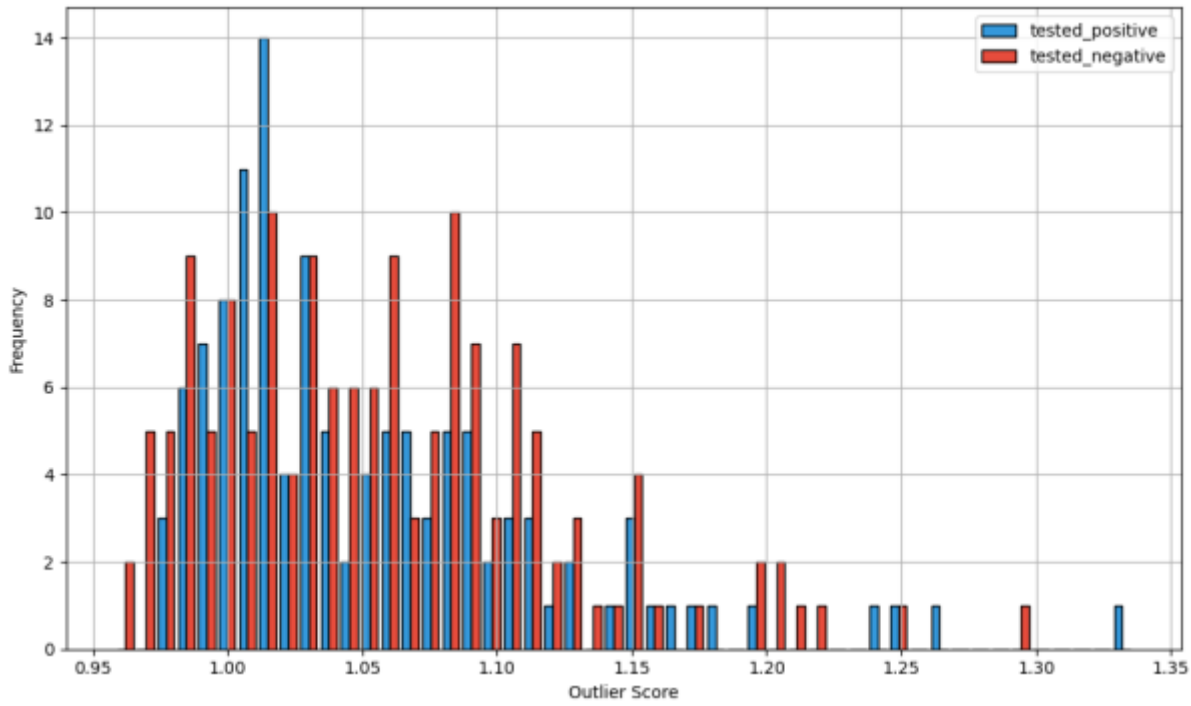


Figure 8: Density based outlier score distribution of HND4

For HND4, the density based outlier detection technique generates outlier scores ranging between 0.96 and 1.33. The resultant histogram for HND4 is shown in figure 8. Among the outlier scores calculated, the top 10 outliers whose outlier score greater than 1.20 are to be identified. For HND4 dataset, the number of instances with higher outlier scores is more when compared to other datasets.

Performance Of Outlier Detection Methods

Machine learning provides several approaches for assessing the performance of learning algorithms. The top ten outliers are removed after using distance and density-based outlier detection algorithms on the HND1, HND2, HND3, and HND4 datasets. After removal, the number of instances in the datasets is reduced to 260 from 270. The performance is then evaluated using a classifier. To compare the performance of various outlier detection techniques, two different classifiers are used. The classifiers are Support Vector Machine (SVM), and Naïve Bayes (NB). These classifiers are employed much in the literature of outlier detection [4].

Evaluation measures analyze various facets of machine learning algorithms. The accuracy measure describes how accurately the set of tuples is classified. Classification accuracy is a popular statistic for evaluating classifiers. The performance of the classifiers is also assessed using the original clean heart disease dataset.

10.48047/jocaaa.2024.33.02.27

Table 3 Performance measure of HND0

Performance measure	Performance (%)	
	NB	SVM
Accuracy	74.3	78.2

Table 3 shows the classification outcomes for the real datasets for the classifiers Naïve Bayes and Support Vector Machine. The accuracy of SVM is significantly higher than that of the Naive Bayes classifier, according to the data (see table 3).

Accuracy of Classifiers

Table 4 shows the accuracy values of classifiers for outlier detection methods for all noisy datasets used. Table 4 shows that the performance of classifiers varies with the outlier detection methods used. Distance-based outlier detection outperforms the other outlier methods for the HND1 dataset. The accuracy of the hybrid SVM and distance-based OD method is higher than that of any other hybrid classifier and outlier detection method, at 78.2%. In general SVM outperforms other models for a 5% heart noise dataset (HND1).

With Naïve bayes as classification method, the density based outlier detection method has greater accuracy of 83.1% for a 10% Heart noise dataset (HND2). It is also observed that, all classifiers employed has its classification accuracy improved when compared to the accuracy obtained before removing outliers for a 10% Heart noise dataset (HND2). HND3 Thus, it is observed from results in table 4 that the compound combination of SVM and density based outlier detection method gives better accuracy 80.9 for 15% Heart noise dataset (HND3). Thus, based on the results in Table 4, it can be concluded that SVM outperforms all other outlier detection methods on all datasets.

10.48047/jocaaa.2024.33.02.27

Table 4 Accuracy of classification methods for all datasets

Method/ Classifier	HND1		HND2		HND3		HND4	
	NB (%)	SVM (%)	NB (%)	SVM (%)	NB (%)	SVM (%)	NB (%)	SVM (%)
Before outlier detection	80.19	85.19	74.07	75.31	83.95	81.48	69.14	72.84
Distance based	74.3	78.2	82.05	80.77	79.49	74.36	76.9	75.64
Density based	75.3	75.3	83.1	79.2	78.2	80.9	72.7	78.9

Conclusion

Outlier detection is a valuable area of research with many real-world applications. Such an algorithm is tailored for the detection of relevant and outlier points in high-dimensional data sets. In this current research, we contrast the performance of some outlier detection algorithms on the heart disease data set. The method's performance is measured for outlier detection by adding varying levels of noise to the original data set. Two classifiers are used for performance measurement. For all the four datasets, SVM classifier is found to generate better performance accuracy. For HND1 dataset, distance based outlier detection method gives more accuracy than density based outlier detection method. But for the remaining three datasets, HND2, HND3, and HND4, density based outlier detection method generates more accuracy. Overall study of accuracy measure for all the four datasets shows that density based outlier detection method performs better than the other outlier detection methods employed.

10.48047/jocaaa.2024.33.02.27

Reference

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Dallas, TX, USA, **2000**, pp. 93–104.

[2] C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for projected clustering of high dimensional data streams," in *Proceedings of the 30th VLDB Conference*, Toronto, Canada, **2004**, pp. 852-863.

[3] S. V. Peshatwar and S. Dongre, "Outlier Detection Over Data Stream Using Cluster Based and Distance Based Approaches," in *Int. Conf. on Electrical Engineering and Computer Science (ICEECS)*, Trivandrum, India, **May 2012**.

[4] Surekha V Peshatwar & Snehlata Dongre, "Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach", *International Conference on Electrical Engineering and Computer Science (ICEECS-2012)*, Trivandrum, **May 12th, 2012**.

[5] V. Mahalakshmi and M. Govindarajan, "Comparison of Outlier Detection Methods in Diabetes Data," *International Journal of Computer Applications*, vol. 155, no. 10, pp. 28–32, **Dec. 2016**.

[6] A. Idri, M. El Asri, and A. Fernández-Alemán, "Systematic Map of Literature Reviews on Clinical Decision Support for Cardiovascular Disease Diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 191, **2020**, 105400.

[7] R. A. Pathan, N. S. Quadri, and N. A. Shaikh, "Impact of Feature Selection Techniques on Heart Disease Prediction Models," *arXiv preprint arXiv:2206.03239*, **2022**.