SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Outlier Detection For Robust Medical Diagnosis: A Benchmark Evaluation Across Classifiers

Bonomali Khuntia[1], Radhanath Patra[2]

[1]Berhampur University, Berhampur, Ganjam Odisha-760 007. Email: bk.cs@buodisha.edu.in
[2]GIET University, Gunupur, Odisha, 765022. Email -1radhanath.patra@gmail.com

Abstract
Outliers in datasets can substantially affect the efficacy of machine learning classifiers, particularly in sensitive domains specificaly healthcare. This study examines multiple outlier detection techniques on benchmark medical datasets, including heart disease, diabetes, and breast cancer. Seven methods spanning statistical, clustering, and proposed approach were applied to each dataset, after which classification was performed independently using Random Forest, Logistic Regression and Support Vector Machine. The results show substantial improvements to the model's resilience and accuracy. Notably, all classifiers achieved 100% accuracy on the heart disease dataset. Random Forest attained 99.07% accuracy on both Wisconsin Breast Cancer Dataset and the Pima Indian Diabetes Dataset.

Keywords: Outlier Detection, Classifier, WBCD, PIDD, CHD, HHD.

1 Introduction

Outliers are noisy data points that degrade machine learning performance, potentially leading to biased models and false predictions [1]. To address this challenge, a variety of statistical and unsupervised outlier detection techniques have been proposed to effectively distinguish well-defined (reliable) data from noisy or anomalous data. These outlier detection techniques are broadly categorized and described as follows statistical method (Z score, Inter Quartile range), Distance Based Method (DBSCAN), Density Based Method (LOF-Local Outlier Factor), Clustering Based Method (K-means), Isolation Based Methods (Isolation Forest) [2].

Standard Deviation Based Outlier Detection

It is a statistical technique refers as "Z-Score" particularly used used for detecting outliers' detection. It is computed as $z=(x-\mu)/\sigma$, where x represents the individual data point μ denotes the meaning of the dataset and σ is the standard deviation [3]. A data point is called as outlier iff it's Z-Score exceeds +3 or falls below -3, indicating that it lies significantly far from the mean of distribution.

Isolation Forest

It applies an unsupervised machine learning algorithm approach to discover anomalous data instances through recursive partitioning [4]. The algorithm performs recursive partitioning by randomly selecting features and corresponding split values, until individual data points are isolated. Each point is assigned an anomaly score, which is then used to determine whether it is an outlier.

Density Based Spatial Clustering Method (DBSCAN)

An unsupervised learning strategy based on clustering principles is employed that identifies dense regions in data, classifies points as core, border and noise, and detects outliers. Clusters are formed by selecting a point randomly and considering two parameters: Epsilon (ε): the neighborhood radius and the MinPoints, which specifies the minimum number of points required within the ε-radius [5]. Based on these, data points are classified into: Core Points: points with at least MinPoints within ε-radius. Border Points: Points with fewer than MinPoints but lying within the ε-radius of a Core Point. Noise: Not part of any cluster. DBSCAN effectively handles multi-dimensional

data and detects outliers as points not belonging to any cluster.

Local Outlier Detection Using LOF

LOF uses k-nearest neighbors to access each data point's local density to identify anomalies. It compares a point's local density with those of its neighbors and point with substantially lower density are identified as outliers [6]. LOF computes: Local Reachability Density (LRD): Local Outlier Factor for each point to assess deviation from neighbors.

K-means Clustering for Outlier detection

K-means Clustering represents a frequently used unsupervised machine learning algorithm for portioning datasets into separate clusters according to similarity among the data points. It is particularly effective for datasets with linearly separable classes. The algorithm is appreciated for its efficiency, robustness, speed, and good computation capability [7,8]. In a typical application, the attributes of the dataset are separated from the class labels, as K-means operates without supervision. The algorithm assigns K centroid initially. Each data instance is associated with the closest centroid using Euclidean distance measure with centroid updates performed iteratively until the convergence is achieved, resulting in K well defined cluster. Points with larger distances are considered less like the cluster and are flagged as potential outliers**.** Typically, a threshold (for example, the top 5% of farthest points) is used to identify outliers. This approach is efficient and effective but works best when clusters are compact and well-separated. Outliers can still affect centroid accuracy.

A proposed K-means Clustering-Based method for Outlier detection

Data grouping and classification represent two distinct machine-learning approaches for data analysis. Clustering organizes data samples into homogenous groups based on intrinsic similarity. In contrast, classification learns the mapping between input attributes and predefined class labels to assign new data instances to specific categories [9]. Here, a combination of K-means clustering algorithm with classification techniques is proposed to remove outliers and hence improve classification performance.

Initially, the features and class labels are separated, and only the feature set is used to perform clustering into two groups. The resulting cluster labels are arbitrary and may not directly match the actual class labels. To eliminate outliers, each cluster is matched to the most common true label within it. This mapping helps align predicted clusters with actual classes. Instances where the assigned cluster label matches the true class label are retained, while those with mismatched labels are identified and treated as outliers. This outlier detection technique is used to extract useful records from anomalous data samples. Eliminating outlier leads to improved classifier performance and benefits classification models adopting this approach. To validate the proposed method, it is applied on the Wisconsin Breast Cancer Dataset (WBCD) [12], the Pima Indians Diabetes Dataset (PIDD) [13], and the Cleveland Heart Dataset (CHD) as well as the Hungarian Heart Dataset (HHD) [14], collected from the UCI Machine Learning Repository of the University of California, Irvine.

Description of Datasets

The datasets utilized in this study are briefly described below.

a. Breast Cancer Dataset

One of the most common and deadly tumors that impact women globally is breast cancer. According to the estimates by the World Health Organization (WHO) in 2025, breast cancer contributed to nearly 15% of female cancer related deaths [15]. Initial stage of breast cancer is frequently curable. So, it is more prominent to properly analyze breast cancer at its initial stage. For analysis, the Wisconsin Breast Cancer Dataset (WBCD) obtained from the UCI Machine Learning Repository is used [16]. It comprises 699 samples with 11 attributes including a class label that is used to categorize tumors as non-cancerous or cancerous. The dataset comprises eleven attributes, including the sample identification number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and a class label (indicating 2 for benign and 4 for cancerous) cases.

b. Diabetes Dataset

Diabetes is a chronic metabolic disorder of global concern, commonly classified as TYPE- I and TYPE-II diabetes mellitus [17]. The rapidly increasing prevalence of diabetes mellitus is associated with a substantial rise in mortality, with approximately 80% of diabetes deaths occurring in developing and underdeveloped countries. So, precise and early prediction is key to managing and controlling diabetes effectively. A widely adopted dataset for diabetes prediction tasks is the PIDD, which originates from the National Institute of Diabetes and Digestive and Kidney diseases [18]. There are nine attributes (eight predicators and one class label). It represents 8 clinical features of 768 women above the age of 21 years. The PIDD falls under supervised binary classification. The Attributes of PIDD dataset include 1) number of pregnancies, 2) the plasma glucose concentration measured two hours after an oral glucose tolerance test, 3) diastolic blood pressure, 4) triceps skinfold thickness, 5) two-hour

serum insulin level, 6) body mass index (BMI), 7) diabetes pedigree function, 8) age, and 9) a class variable indicating diabetic or non-diabetic status. Among the 768 records, 500 belong to non-diabetic patients, and 268 to diabetic patients. The dataset exhibits an uneven class distribution, making data analysis particularly challenging. Therefore, using proper machine-learning techniques is essential for accurate and reliable prediction.

c. Heart Dataset

Heart disease remains to be the world's greatest cause of martality for both men and woman. Heart failure, heart stroke, congenital heart disease, and coronary artery disease are major causes of causalities and morbidity. According to recent reports, heart diseases account for approximately 31% of total death globally [19]. Given the growing prevalence, and severity of this condition, there is an urgent need to implement effective machine learning algorithims for timely detection and precise prediction. In the analysis, the Cleveland and the Hungarian Heart Disease datasets are utilized. The CHD contains 303 instances and 76 attributes. In the study, the refined version consisting of 14 attributes (13 input features and one output class label) [20] is employed. These attributes include 1) Age: person's age in years, 2) Sex, 3) Type of Chest Pain (Cp), 4) Blood Pressure at Rest(Trestbps) measured in mmHg, 5) Serum Cholesterol(Chol) measurement in mg/dl, 6) Fasting Blood Sugar (Fbs), 7) Resting ECG results(Restecg), 8) Maximum Heart Rate Achieved (Thalach), 9) Exercise Induced Angina (Exang), 10) ST Depression (Oldpeak), 11) Slope of Peak Exercise ST Segment (Slope), 12) Number of Major Vessels Colored by Fluoroscopy (Ca), 13) Thalassemia status (Thal), and 14) Num the class label denoting the presence(1) or absence (0) of heart disease.

2 Related Work

A lot of research has been done on disease classification and outlier identification using advanced machine learning (ML) techniques across various healthcare datasets, with a particular emphasis on the Pima Indian Diabetes Dataset (PIDD). Aldhyani et al. [21] introduced a Rough K-Means (RKM) clustering framework to classify lower approximation data, achieving an accuracy of 80.55% using a Naïve Bayes classifier. Torkey et al. [22] improved classification performance by applying cluster-mean-based imputation along with deep learning and Random Forest, yielding 94.5% accuracy. Similarly, Li et al. [23] developed a Harmony K-Means clustering approach integrated with K-Nearest Neighbors (KNN), reporting 91.65% accuracy with nine selected features. Ahamed et al. [24] demonstrated the effectiveness of data preprocessing and Light Gradient Boosting Machine (LGBM), attaining 95.20% accuracy. Annamalai et al. [25] presented a hybrid framework combining the SMVMF framework with an Optimized Weighted Deep Artificial Neural Network (OWDANN), also achieving 95.20%. More recently, Nnamoko [26] integrated IQR-based outlier replacement, SMOTE for class balancing, and C4.5 classification, reaching 89.5% accuracy. Dashdondov et al. [27] utilized Mahalanobis Distance for outlier removal and feature selection with XGBoost, attaining 98.4%. Joglekar et al. [28] adressed a skewness-based outlier detection algorithim that enhanced logistic regression accuracy to 84%. The most recent and advanced contribution by Farnoosh et al. [29] applied Kernel Density Estimation (KDE) with K-Medoids and Gaussian Naïve Bayes, and a parallel KDE-based K-Means model, achieving a remarkable 98.6% accuracy on a Type 2 Diabetes dataset.

For the Heart Disease Dataset, Sumwiza et al. [30] applied Interquartile Range (IQR) outlier detection and correlation driven feature selection integrating with a Random Forest classifier, achieving an accuracy of 86% on 1025 samples. Kukala et al. [31] employed Z-score-based outlier detection, Grey Wolf Optimization (GWO) for feature selection, and AdaBoost classification to achieve 89% accuracy. Ripan et al. [32] used a silhouette-guided K-Means clustering approach to eliminate irrelevant clusters, followed by Random Forest, achieving 88%. Hasan et al. [33] applied a robust pipeline incorporating SMOTE-Tomek resampling, feature selection, and outlier detection before training a Random Forest model, achieving an impressive 96%. Shah et al. [34] attained 90.78% accuracy using preprocessing and KNN (k=7) on 13 attributes. Rani et al. [35] integrated MICE imputation, GARFE feature selection, and Random Forest, yielding 86.60% accuracy on an 8-feature subset. Gokulnath et al. [36] combined Genetic Algorithm-based feature selection with SVM to reach 88.4% accuracy using just 7 features.

Concerning the Wisconsin Breast Cancer Dataset (WBCD/WDBC), various hybrid models have shown high performance. Zheng et al. [37] proposed a hybrid K-Means-SVM classifier incorporating membership-function-based feature optimization, attaining 97.38% accuracy. Wang et al. [38] Introduced a dynamically optimized K-Means clustering method for simultaneous clustering and data imputation, reaching 96.67%. Mohamed and Farouk [39] designed a Geometrical-Driven Diagnosis (GDD) classifier using modified K-Means and feature selection, reporting 95.8% accuracy. Aldhyani et al. [40] again applied RKM clustering to filter ambiguous data before SVM classification, achieving 97.53%. Ali and Shehata [41] introduced a cohort-based outlier rejection technique and information gain-based feature selection followed by an Extreme Learning Machine (ELM), reaching 98.7% accuracy. Alhayali et al. [42] used an ensemble of Hoeffding Tree and Naïve Bayes, obtaining 95.99%. Roguia and Mohamed [43] employed Particle Swarm Optimization (PSO) with K-Means and a Backpropagation Neural Network to achieve 97.82%. Karthik et al. [44] utilized Recursive Feature Elimination (RFE) with a Deep Neural

Network (DNN), reporting the highest accuracy of 98.62%. Most recently, Tintu et al. [45] proposed a hybrid outlier detection model (HOTSM) for WDBC, combining IQR and Isolation Forest with Pearson correlation-based dimensionality reduction, and achieved 98.4% accuracy using a Random Forest classifier.
Authors in [21]-[23] emphasized using modified K-Means clustering to choose the optimal attributes for prediction. However, in [24]-[26] authors highlighted the significance of calculating the cluster centroid using an objective function for clustering. In [27], KNN clustering depends on the suitable selection of values of K and may not perform well for higher values of data. Authors in [28]-[45] emphasized on feature selection for accuracy improvement.
In the present work, a novel methodology leveraging K-Means clustering is proposed for the selection of informative records. The study primarily aims to improve the classification performance on structured medical datasets. The contents of the paper are fourfold (i) Preprocessing-handling missing values to ensure clean input data, (ii) K-Means Clustering [46] is applied to partition the dataset into two groups for clustering, and assigning class labels based on majority class within each cluster, (iii) Outlier Detection: identifying relevant records by comparing the labels assigned through clustering with the original dataset class labels to filter inconsistent or noisy data, (iv) Classification: evaluating the effectiveness of various weak classifier on the refined dataset to improve predictive accuracy.

3 Result Analysis

Classification Performance on Heart Dataset:

The dataset contains 1025 samples. The classification performance was accessed through the application of Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression Models. Table 1 presents performance metrics of these classifiers without any outlier detection method.

Table 1: Classification Performance Without Outlier Detection

Dataset	Classifier	Accuracy	Precision	Recall	F-Score
Heart Dataset (1025 Samples)	Support Vector Machine	0.88	0.89	0.88	0.88
	Logistic Regression	0.79	0.80	0.79	0.79
	Random Forest	0.98	0.98	0.98	0.98

Subsequently, various outlier detection methods are applied to the heart disease dataset to enhance classification performance. The resulting sample sizes after applying each method are as follows: K-Means clustering retained 822 samples, Z-score retained 969 samples, Isolation Forest retained 973 samples, DBSCAN retained 499 samples, and Local Outlier Factor retained 974 samples. These refined datasets are used to train and evaluate LR, SVM and RF classifiers. The results are shown in Table 2 with a visual representation provided in the accuracy heat map shown in Figure 1. The table presents the performance of each classifier across different outlier detection methods with metrics such as accuracy, precision, recall and F-score.

Table 2: Classification Performance with Outlier Detection Methods

Dataset	Outlier Method	Samples	Classifier	Accuracy	Precision	Recall	F-Score
Heart Dataset	K-means clustering	822	Support Vector Machine	1	1	1	1
			Logistic Regression	1	1	1	1
			Random Forest	1	1	1	1
	Z SCORE	969	Support Vector Machine	0.92	0.92	0.92	0.92
			Logistic Regresion	0.88	0.88	0.88	0.88
			Random Forest	1	1	1	1

	Isolation Forest	973	Support Vector Machine	0.96	0.97	0.96	0.96
			Logistic Regression	0.83	0.84	0.83	0.83
			Random Forest	1	1	1	1
	DBSCAN	499	Support Vector Machine	0.93	0.93	0.93	0.93
			Logistic Regression	0.94	0.93	0.94	0.93
			Random Forest	1	1	1	1
	Local Outlier Factor	974	Support Vector Machine	0.91	0.91	0.91	0.91
			Logistic Regression	0.75	0.76	0.75	0.75
			Random Forest	0.72	0.72	0.72	0.72

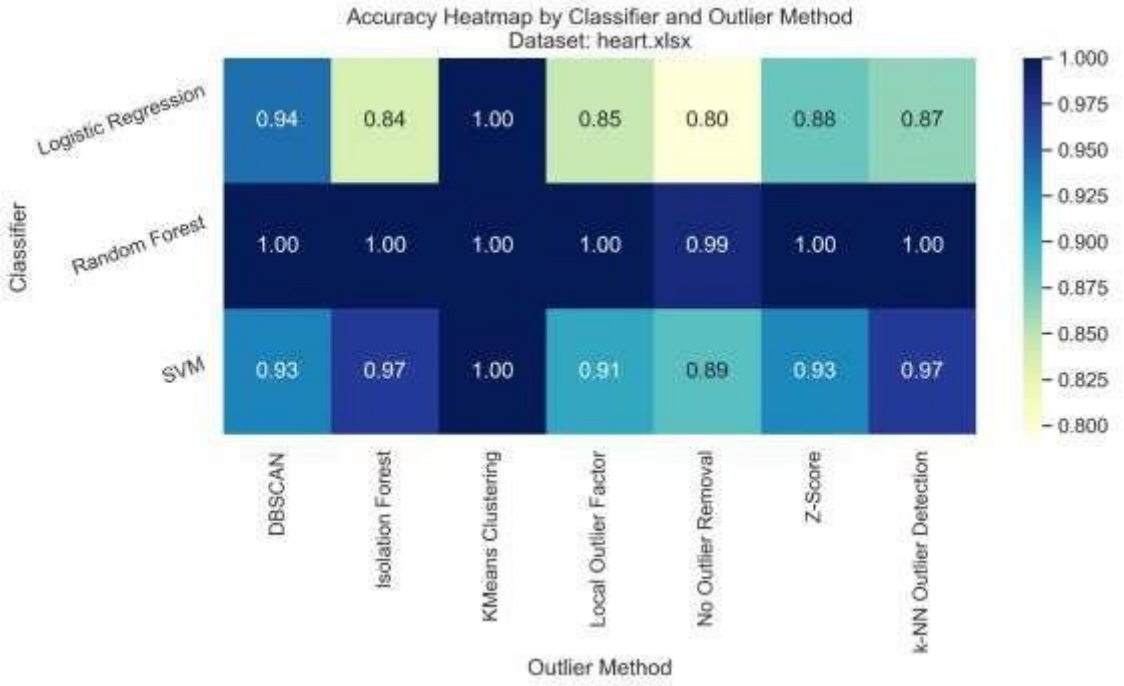

Fig 1: Accuracy Heatmap for the Heart dataset

Classification Performance of PIDD:
The PIDD contains 768 samples. Three commonly used machine learning classifiers namely SVM, LR, and RF are used. Table 3 shows the classifier performance without any outlier detection method.

Table 3: Classification Performance without any Outlier Detection Method

Dataset	Classifier	Accuracy	Precision	Recall	F-Score
Diabetes Dataset (768 Samples)	Support Vector Machine	0.72	0.72	0.72	0.72

	Logistic Regression	0.75	0.75	0.75	0.75
	Random Forest	0.72	0.72	0.72	0.72

Outlier detection methods were applied to the PIDD to enhance classification performance. The number of samples retained after applying each method were as follows: K-Means clustering retained 518 samples, Z-Score retained 688 samples, Isolation Forest retained 729 samples, DBSCAN retained 733 samples, and Local Outlier Factor retained 729 samples. These refined datasets were subsequently used for classification using SVM, LR and RF classifiers as shown the in Table 4 the performance parameters, and with a visual representation provided in the accuracy heat map shown in Figure 2.

Table 4: Classification Performance with Outlier Detection Methods

Dataset	Outlier Method	Samples	Classifier	Accuracy	Precision	Recall	F-Score
Diabetes Dataset	K-means clustering	518	Support Vector Machine	0.96	0.96	0.96	0.96
			Logistic Regression	0.98	0.98	0.98	0.98
			Random Forest	0.99	0.99	0.99	0.99
	Z-Score	688	Support Vector Machine	0.76	0.77	0.76	0.75
			Logistic Regression	0.77	0.79	0.77	0.75
			Random Forest	0.73	0.72	0.73	0.72
	Isolation Forest	729	Support Vector Machine	0.76	0.75	0.76	0.75
			Logistic Regression	0.76	0.75	0.76	0.76
			Random Forest	0.78	0.77	0.78	0.77
	DBSCAN	733	Support Vector Machine	0.77	0.77	0.77	0.76
			Logistic Regression	0.78	0.78	0.78	0.76
			Random Forest	0.74	0.73	0.74	0.72
	Local Outlier Factor	729	Support Vector	0.76	0.76	0.76	0.76
			Logistic Regression	0.79	0.79	0.79	0.79
			Random Forest	0.78	0.78	0.78	0.78

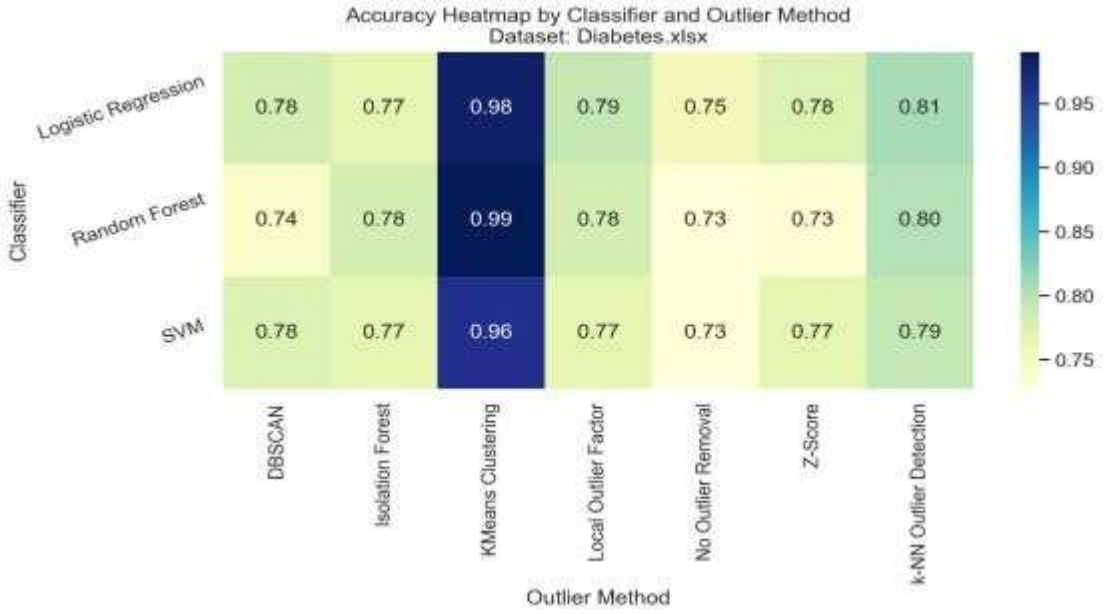

Fig 2: Accuracy Heatmap for the Diabetes Dataset

Classification Performance of Wisconsin Breast Cancer Dataset (WBCD):
There are 683 samples in the breast cancer dataset. Classification performance was evaluated using SVM, LR and RF. Initially, these classifiers were trained and tested on the dataset without incorporating any outlier detection techniques. The performance of the classifiers without any outlier detection technique are presented in Table 5.

Table 5: Classification Performance without any Outlier Detection Method

Dataset	Classifier	Accuracy	Precision	Recall	F-Score
Breast Cancer Dataset (683 Samples)	Support Vector Machine	0.97	0.97	0.97	0.97
	Logistic Regression	0.95	0.95	0.95	0.95
	Random Forest	0.96	0.96	0.96	0.96

Outlier detection methods were applied to the above dataset to enhance classification performance. After applying each method, the number of retained samples varied: K-Means clustering resulted in 651 samples, Z-Score retained 632 samples, Isolation Forest retained 648 samples, Density based spatial clustering of Applications with Noise (DBSCAN) retained 625 samples, and Local Outlier Factor retained 648 samples. These refined datasets are used to perform classification using SVM, LR, and RF classifiers. The corresponding performance outcomes are summarized in Table 6. The impact of outlier detection on classifier performance is illustrated through various visualizations, with accuracy heatmap shown in Figure 3, the classification accuracy presented in Figure 4, as well as the F1-score in Figure 5, and the average accuracy in Figure 6.

Table 6: Classification Performance with Outlier Detection Methods

Dataset	Outlier Method	Samples	Classifier	Accuracy	Precision	Recall	F-Score
Breast Cancer Dataset	K-means clustering	651	Support Vector Machine	0.97	0.97	0.97	0.97
			Logistic Regression	0.99	0.99	0.99	0.99
			Random Forest	0.99	0.99	0.99	0.99

	Z SCORE	632	Support Vector Machine	0.99	0.99	0.99	0.99
			Logistic Regression	0.99	0.99	0.99	0.99
			Random Forest	0.99	0.99	0.99	0.99
	Isolation Forest	648	Support Vector Machine	0.96	0.96	0.96	0.96
			Logistic Regression	0.96	0.97	0.96	0.96
			Random Forest	0.94	0.94	0.94	0.94
	DBSCAN	625	Support Vector Machine	0.99	0.99	0.99	0.99
			Logistic Regression	0.99	0.99	0.99	0.99
			Random Forest	0.99	0.99	0.99	0.99
	Local Outlier Factor	648	Support Vector Machine	0.96	0.96	0.96	0.96
			Logistic Regression	0.96	0.96	0.96	0.96
			Random Forest	0.96	0.96	0.96	0.96

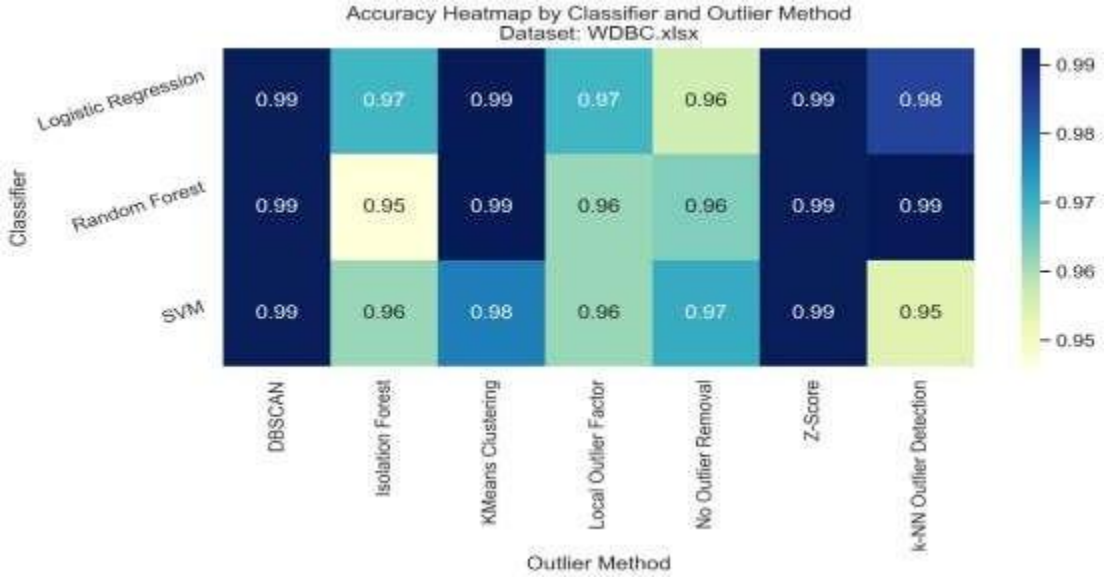

Fig 3: Accuracy Heatmap for the WDBC Dataset

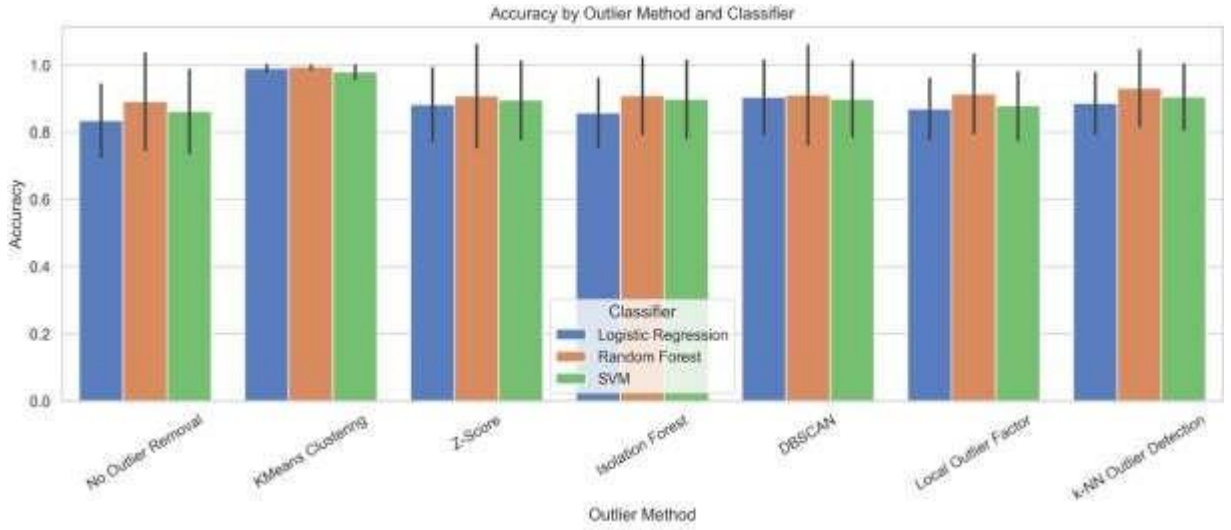

Fig 4: Accuracy

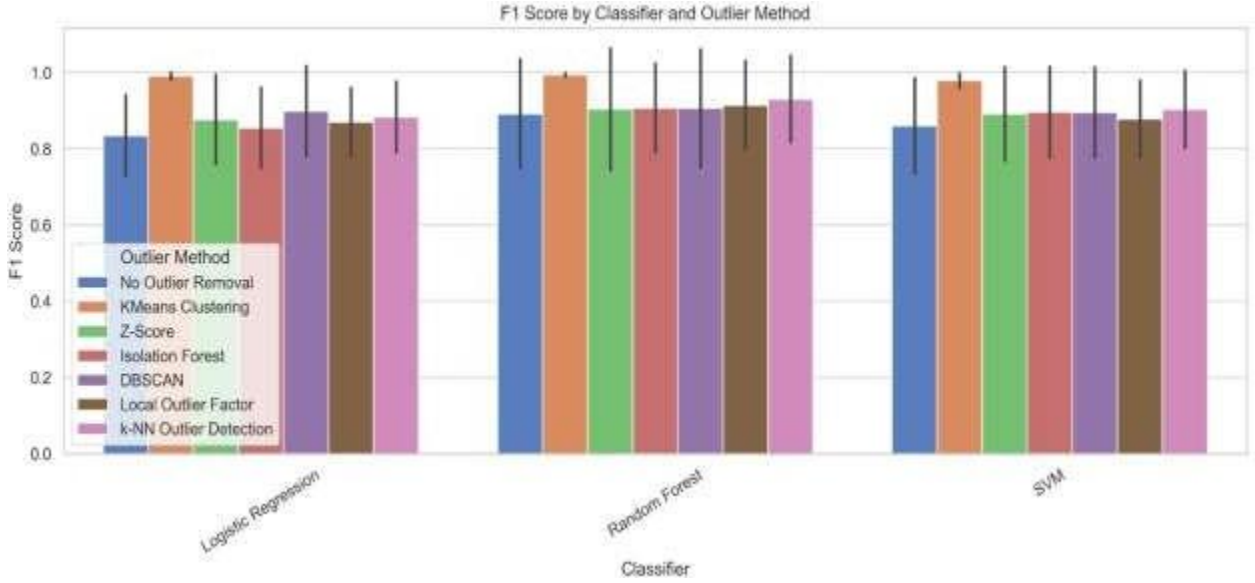

Fig 5: F1-Score

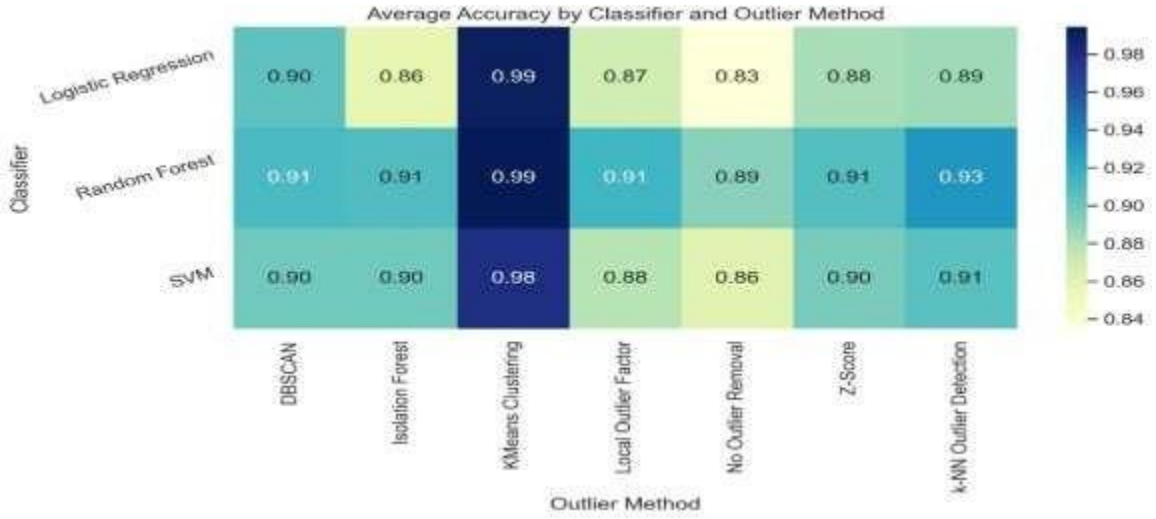

Fig 6: Average Accuracy

Assessment Related to Existing Approaches

Table 12: Accuracy Comparison of the proposed method and existing CHD and HHD studies

Dataset	Methods	Authors	Accuracy
CHD with 13 attributes	Chi-square and Random Forest	Karthick et al. [44]	88.5%
	Jellyfish Optimization and SVM	Ahmad and Pollat [47]	98.09%
	Z Score, SMOTE, Correlation and Random Forest	W Wang. [48]	98.50%
	Ensemble method based on Random Forest	Sumwiza et al. [49]	99.00%
	HRFLM- Hybrid Random Forest with Linear Model	Senthil Kumar Mohan et al. [50]	88.70%
	SVM and $\chi 2$	Sarra et al. [51]	89.47%
	Proposed Technique		**100%**

Table 13: **Accuracy Comparison of the proposed method and existing PIDD studies**

Dataset	Methods	Authors	Accuracy
PIDD with 8 attributes	Linear SVM	Kaur, H., & Kumari, V [52]	89%
	Ensemble method	Alehegn, M., Joshi, R., &Mulay, P [53]	93.62%
	Gradient boosting	Birjais, R et al. [54]	86%
	Preprocessing and KNN classifier	Kumari et al. [55]	92.28%
	PCA + Logistic Regression model	Roopa et al. [56]	82.1%
	PCA+k-means+Logistic Regression	Zhu, Cetal. [57]	97.4%
	Proposed Technique		**99.13%**

Table 14: Accuracy Comparison of the proposed method and existing WBCD studies

Dataset	Methods	Authors	Accuracy
WBCD with 9 attributes	Naïve bayes, RBF, J48	Vikas Chaurasia et al. [58]	97.3%
	WDBC with tabu feature selection+KNN	Habib Dhahri et al. [59]	98.24%
	SVM+AR	Abderrahmane Ed-daoudy et al. [60]	98%
	K boosted C5	Jue zhang et al. [61]	97.5%
	MLP	Ali Al bataineh [62]	99%
	XGBooSt+RFE	Abdulkareem et al. [63]	99.02%
	REF-SVM	Muhammad hammad et al. [64]	99%
	Naïve bayes, RBF, J48	Vikas Chaurasia et al. [58]	97.3%
	Proposed Technique		99.00%

4 Conclusion

This work focuses on the selection of relevant records from a dataset by discarding outliers, which is used by different classifiers to improve classification performance. Initially, the proposed technique is applied to the Cleveland Heart Dataset with different classifiers. The result shows satisfactory improvement in classification accuracy. Subsequently, it is also validated with other datasets such as Breast cancer, Pima Indian diabetes dataset, Indian liver patient dataset, etc. The result demonstrate a significant improvement in precision, recall, F-measure and ROC metrics for all evaluated classifiers and datasets. It can also be used along with feature selection which needs further study. This technique can also be used with image datasets for improving classification performance.

References:

1. Falahi, T., Nassreddine, G., & Younis, J. (2023). Detecting Data Outliers with Machine Learning. Al-Salam Journal for Engineering and Technology, 152-164.
2. Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. ACM Computing Surveys (CSUR), 53(3), 1-37.
3. Abdi, H. (2007). Z-scores. Encyclopedia of measurement and statistics, 3, 1055-1058.
4. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.
5. Bi, F. M., Wang, W. K., & Chen, L. (2012). DBSCAN: density-based spatial clustering of applications with noise. J. Nanjing Univ, 48(4), 491-498.
6. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).
7. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.
8. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.
9. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.
10. B. Kulis and M. I. Jordan, "Revisiting k-means: new algorithms via Bayesian nonparametrics," in Proceedings of the 29th International Conference on Machine Learning (ICML '12), pp. 513–520, Edinburgh, UK, July 2012.
11. M. S. Yang and K. P. Sinaga, "A feature-reduction multi-view k-means clustering algorithm," IEEE Access, vol. 9, p. 1, 2019.
12. UCI Machine Learning Repository. Breast Cancer Wisconsin (Original) Data Set: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
13. Indian Pima Indian Diabetes Dataset. Available Online: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed on 20th March,2023)
14. UCI Machine Learning Repository. https//archive.ics.uci.edu/ml/datasets/Heart+Disease.
15. Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer medicine, 9(9), 3234-3243.
16. Iqbal, M. S., Ahmad, W., Alizadehsani, R., Hussain, S., & Rehman, R. (2022, November). Breast cancer dataset, classification and detection using deep learning. In Healthcare (Vol. 10, No. 12, p. 2395). MDPI.

17. Yabo, M. M. I., Garko, A. B., Muslim, A. A., & Suru, H. U. (2022). A review of diabetes datasets. Journal of Computer Sciences and Applications, 10(1), 6-15.
18. Asha, V., Lamani, M. R., Padmaja, K., & Gondhale, T. A. (2024). Enhancing Diabetes Prediction Accuracy with AdvanSVM: A Machine Learning Approach Using the PIMA Dataset. Seybold Report Journal, 19(05), 51-72.
19. Gaidai, O., Cao, Y., & Loginov, S. (2023). Global cardiovascular diseases death rate prediction. Current problems in cardiology, 48(5), 101622.
20. Govindaraju, K., & Kalimuthu, G. (2024, November). Heart Disease Analysis and Prediction with Machine Learning Techniques Using Cleveland Dataset. In International Symposium on Intelligent Computing Systems (pp. 20-43). Cham: Springer Nature Switzerland.
21. Aldhyani, T. H., Alshebami, A. S., &Alzahrani, M. Y. (2020). Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. Journal of healthcare engineering, 2020.
22. Torkey, H., Ibrahim, E., Hemdan, E. E. D., El-Sayed, A., & Shouman, M. A. (2022). Diabetes classification application with efficient missing and outliers data handling algorithms. Complex & Intelligent Systems, 1-17.
23. Li, X., Zhang, J., & Safara, F. (2021). Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural Processing Letters, 1-17.
24. Ahamed, B. S., Arya, M. S., & Nancy V, A. O. (2022). Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques. Frontiers in Computer Science, 4, 835242.
25. Annamalai, R., & Nedunchelian, R. (2021). Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. Computational Intelligence and Neuroscience, 2021.
26. Nnamoko, N., & Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction. Artificial intelligence in medicine, 104, 101815.
27. Dashdondov, K., Lee, S., & Erdenebat, M. U. (2024). Enhancing Diabetes Prediction and Prevention through Mahalanobis Distance and Machine Learning Integration. Applied Sciences, 14(17), 7480.
28. Joglekar, P., Katale, T., Deshpande, S., Kelgandre, S., Katale, A., & Chotrani, A. (2024). Skewness based Outlier Elimination Algorithm for Pima Diabetes Dataset. Grenze International Journal of Engineering & Technology (GIJET), 10, 12.
29. Farnoosh, R., Abnoosian, K., & Isewid, R. A. (2025). Two machine-learning hybrid models for predicting type 2 diabetes mellitus. Journal of Medical Signals & Sensors, 15(4), 11.
30. Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked, 41, 101316.
31. Kukkala, V. R., Praveen, S. P., Tirumanadham, N. S. K. M. K., & Srinivasu, P. N. (2024). A Study on Outlier Detection and Feature Engineering Strategies in Machine Learning for Heart Disease Prediction. Computer Systems Science & Engineering, 48(5).
32. Ripan, R. C., Sarker, I. H., Hossain, S. M. M., Anwar, M. M., Nowrozy, R., Hoque, M. M., & Furhad, M. H. (2021). A data-driven heart disease prediction model through K-means clustering-based anomaly detection. SN Computer Science, 2(2), 112.
33. Hasan, M., Sahid, M. A., Uddin, M. P., Marjan, M. A., Kadry, S., & Kim, J. (2024). Performance discrepancy mitigation in heart disease prediction for multisensory inter-datasets. PeerJ Computer Science, 10, e1917.
34. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1, 1-6.
35. Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), 263-275.
36. Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22, 14777-14787.
37. Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.
38. Wang, S., Li, M., Hu, N., Zhu, E., Hu, J., Liu, X., & Yin, J. (2019). K-means clustering with incomplete data. IEEE Access, 7, 69162-69171.
39. Mohamed, A. E., & Farouk, M. (2020). GDD: Geometrical driven diagnosis based on biomedical data. Egyptian Informatics Journal, 21(3), 183-190.
40. Aldhyani, T. H., Alshebami, A. S., & Alzahrani, M. Y. (2020). Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. Journal of healthcare engineering, 2020(1), 4984967.
41. Ali, S. H., & Shehata, M. (2024). A New Breast Cancer Discovery Strategy: A Combined Outlier Rejection Technique and an Ensemble Classification Method. Bioengineering, 11(11), 1148.
42. Alhayali, R. I., Ahmed, M. A., Mohialden, Y. M., & Ali, A. H. (2020). Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes. Indonesian Journal of Electrical Engineering and Computer Science, 18(2), 1074-1080.

43. Roguia, S., & Mohamed, N. (2019). An optimized RBF-neural network for breast cancer classification. International Journal of Informatics and Applied Mathematics, 1(1), 24-34.
44. Karthik, S., Srinivasa Perumal, R., & Chandra Mouli, P. V. S. S. R. (2018). Breast cancer classification using deep neural networks. Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1, 227-241.
45. Tintu, P. B., Veni, S., & Priya, S. M. (2024). An Effective Hybrid Outlier Selection Method for Breast Cancer Classification Using Machine Learning Algorithms. Indian Journal of Science and Technology, 17(43), 4494-4501.
46. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.
47. Ahmad, A. A., & Polat, H. (2023). Prediction of heart disease based on machine learning using jellyfish optimization algorithm. Diagnostics, 13(14), 2392.
48. Wang, W. (2024). Heart disease prediction using machine learning models. Theoretical and Natural Science, 51, 9-17.
49. Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked, 41, 101316.
50. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.
51. Sarra, R. R., Dinar, A. M., Mohammed, M. A., & Abdulkareem, K. H. (2022). Enhanced heart disease prediction based on machine learning and $\chi 2$ statistical optimal feature selection model. Designs, 6(5), 87.
52. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics.
53. Alehegn, M., Joshi, R. R., &Mulay, P. (2019). Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes And J48: An Ensemble Approach. Int. J. Sci. Technol. Res, 8(9), 1346-1354.
54. Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), 1-8.
55. kumari, M., & Ahlawat, P. (2021). DCPM: An effective and robust approach for diabetes classification and prediction. International Journal of Information Technology, 13, 1079- 1088.
56. Roopa, H., & Asha, T. (2019). A linear model based on principal component analysis for disease prediction. IEEE Access, 7, 105314-105318.
57. Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179.
58. Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 12(2), 119-126.
59. Dhahri, H., Rahmany, I., Mahmood, A., Al Maghayreh, E., &Elkilani, W. (2020). Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions. BioMed Research International, 2020.
60. Ed-daoudy, A., &Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. Network Modeling Analysis in Health Informatics and Bioinformatics, 9, 1-10.
61. Zhang, J., Chen, L., & Abid, F. (2019). Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method. Journal of healthcare engineering, 2019.
62. Bataineh, A. A. (2019). A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. International Journal of Machine Learning and Computing, 9(3), 248-254.
63. Abdulkareem, S. A., & Abdulkareem, Z. O. (2021). An evaluation of the Wisconsin breast cancer dataset using ensemble classifiers and RFE feature selection. Int. J. Sci., Basic Appl. Res., 55(2), 67-80..
64. Memon, M. H., Li, J. P., Haq, A. U., Memon, M. H., & Zhou, W. (2019). Breast cancer detection in the IOT health environment using modified recursive feature selection. wireless communications and mobile computing, 2019.