Volume 13, Issue 1, Page 70-77, 2024 https://doi.org/10.46810/tdfd.1364397 Research Article

BİNGÖL UNIVERSITY

Türk Doğa ve Fen Dergisi
Turkish Journal of Nature and Science
www.dergipark.gov.tr/tdfd

An Investigation on the Use of Clustering Algorithms for Data Preprocessing in Breast Cancer Diagnosis

Ali ŞENOL[1*], Mahmut KAYA[2]

[1] Tarsus University, Engineering Faculty, Department of Computer Engineering, Mersin, Türkiye
[2] Firat University, Engineering Faculty, Department of Artificial Intelligence and Data Engineering, Elazig, Türkiye

Ali ŞENOL ORCID No: 0000-0003-0364-2837
Mahmut KAYA ORCID No: 0000-0002-7846-1769

**Corresponding author: alisenol@tarsus.edu.tr*

(Received: 21.09.2023, Accepted: 26.02.2024, Online Publication: 26.03.2024)

Keywords
Outlier detection, Clustering, Classification, Breast cancer diagnosis.

Abstract: Classification algorithms are commonly used as a decision support system for diagnosing various diseases, such as breast cancer. However, the accuracy of classification algorithms can be affected negatively if the data contains outliers and/or noisy data. For this reason, outlier detection methods are frequently used in this field. In this study, we propose and compare various models that use various clustering algorithms to detect outliers in the data preprocessing stage of classification to investigate their effects on classification accuracy. Clustering algorithms such as DBSCAN, HDBSCAN, OPTICS, FuzzyCMeans, and MCMSTClustering (MCMST) were used separately in the data preprocessing stage of the k Nearest Neighbor (kNN) classification algorithm for outlier elimination, and then the results were compared. According to the results, the kNN + MCMST model most effectively eliminated outliers. The classification accuracy of the kNN + MCMST model was 0.9834, which was the best one, while the accuracy of the kNN algorithm without using any data preprocessing was 0.9719.

Meme Kanseri Teşhisinde Kümeleme Algoritmalarının Veri Ön İşleme Amacıyla Kullanılması Üzerine Bir İnceleme

Anahtar Kelimeler
Sapan veri tespiti, Kümeleme, Sınıflandırma, Meme kanseri teşhisi.

Öz: Sınıflandırma, meme kanseri teşhisinde olduğu gibi pek çok hastalığın teşhisi konusunda karar destek sistemleri olarak kullanılmaktadır. Verilerin sapan ve/veya gürültülü veri içermesi durumunda sınıflandırma algoritmalarının başarısı olumsuz etkilenebilmektedir. Bu nedenle bu alanda sapan veri tespit yöntemleri sıkça kullanılmaktadır. Bu çalışmada sapan verileri tespit etmek amacıyla çeşitli kümeleme algoritmalarının sınıflandırmanın veri ön işleme aşamasında kullanılması durumunda sınıflandırma başarısının nasıl etkileneceğine yönelik modeller önerilmekte ve kıyaslanmaktadır. Kümeleme algoritmalarından DBSCAN, HDBSCAN, OPTICS, FuzzyCMeans ve MCMSTClustering (MCMST) algoritmaları k en yakın komşu (kNN) sınıflandırma algoritmasının veri ön işleme aşamasında sapan verileri ortadan kaldırma amacıyla ayrı ayrı kullanılmış ve sonuçlar karşılaştırılmıştır. Elde edilen sonuçlara göre MCMST algoritmasının sapan verileri ortadan kaldırmada daha başarılı olduğu tespit edilmiştir. Veri önişleme işlemi yapılmaksızın kNN algoritmasının kullanılması durumunda sınıflandırma başarısı 0.9719 iken; en yüksek sınıflandırma başarısına ulaşan kNN + MCMST modelinin doğruluk oranının 0.9834 olduğu tespit edilmiştir.

1. INTRODUCTION

Breast cancer is the most common type of cancer among women and has a high mortality rate if not diagnosed and treated in time. Every year, 2 million 800 thousand women worldwide are diagnosed with breast cancer. However, 90% of patients successfully overcome breast cancer with early diagnosis and treatment [1]. This underscores the critical importance of early detection strategies in combating breast cancer and reducing its mortality rates.

Classification, a sub-branch of machine learning, is used in many areas [2-6]. One of these areas is health

applications. The use of classification algorithms as a decision support system contributes significantly to the diagnosis of diseases. Several classification algorithms are commonly employed in healthcare settings for disease diagnosis, including breast cancer. Because, classification algorithms can predict new arrival data by learning from existing ones. Naïve Bayes [7], Support Vector Machines (SVM) [8], kNN [9], Decision Trees [10], and Artificial Neural Networks [11] are widely used in this field. These algorithms demonstrate varying degrees of effectiveness in accurately classifying medical data and assisting healthcare professionals in making informed decisions regarding patient care.

The most important factor that reduces the classification accuracy of classification algorithms in machine learning is the presence of outliers and noisy data. Outlier data can arise due to data processing errors, sampling errors, data entry errors, and natural causes (changes in the data). This kind of data can cause classification algorithms to learn the data incorrectly, thus reducing their accuracy. Data preprocessing serves as a crucial step in enhancing the robustness of classification algorithms against outliers. Numerous methods have been proposed in the literature to detect outliers. Isolated Forests [12], Local Outlier Factor (LOF) [13] One-Class SVM [14], and IQR [15, 16] are the leading methods of this area. In addition, clustering algorithms are also frequently used to detect outliers. K-means [17] and DBSCAN [18] are two of these algorithms. Clustering algorithms are used to detect outliers as they can assume that data outside clusters are outliers.

Efficient outlier detection holds paramount importance in classification tasks, particularly in healthcare applications such as breast cancer diagnosis. Identifying and mitigating outliers not only improves the accuracy of classification algorithms but also enhances the reliability of diagnostic decisions. By leveraging advanced outlier detection techniques, healthcare professionals can ensure that classification models are trained on high-quality, representative data, leading to more precise and actionable insights. Thus, the integration of robust outlier detection methodologies into the data preprocessing pipeline is essential for optimizing the performance of classification algorithms and ultimately improving patient outcomes.

In this study, DBSCAN [18], HDBSCAN [19], OPTICS [20], FuzzyCMeans [21], and MCMST [22] clustering algorithms were used to reveal their contribution to the success of classification algorithms when they are used to detect outliers in the data preprocessing stage of classification. To reveal the performance of the models, the obtained results were compared in terms of both classification success and run-time complexity. So, the main contribution of this study to the literature can be summarized as follows:

- Different clustering algorithms were used in the data preprocessing stage of classification and their contribution to classification accuracy was analyzed.
- The MCMST algorithm was used for the first time in this study to detect outliers in data preprocessing and contributed significantly to high classification accuracy.

The rest of the paper is organized as follows: The second section discusses the literature review, while the third section provides information about the algorithms used in this paper. Next, section four presents detailed information about the proposed models. Then, in the fifth section, we provide details about the experimental study and setup. In the sixth section, we share and discuss the results. Finally, in the seventh section, we conclude the study and share plans for future works.

2. LITERATURE REVIEW

In recent years, numerous studies have explored the application of machine learning algorithms, including ANNs , SVM, Naïve Bayes, and kNN, for breast cancer diagnosis using the Wisconsin Breast Cancer Dataset (WBCD). While these studies have reported high classification accuracies, a critical examination reveals certain drawbacks and gaps that warrant further investigation.

One of these studies was proposed by Chen et al. in [23] in 2011. The authors aimed to diagnose breast cancer using rough sets and SVM. Their proposed model achieved 99.41% classification accuracy. Marcano-Cedeno and Andina [24] used ANN with metaplasticity-based multilayer perceptron algorithm for breast cancer diagnosis and achieved 96.26% accuracy. In another study, Seera and Lim [25], proposed an intelligent system for breast cancer diagnosis with a hybrid model including a Fuzzy Min-Max Neural Network, Regression Tree, and Random Forest algorithms. It was found that the model they proposed reached 98.84% classification accuracy. In the proposed model, the Fuzzy Min-Max Neural Network was responsible for incremental learning, the Regression Tree for data intelligibility, and the Random Forest for improving prediction accuracy. Another study in this field was carried out by Zheng et al. [26] to classify breast cancer dataset using SVM with k-means clustering algorithm. In their proposed model, the k-means algorithm clusters the data into cancerous and non-cancerous clusters, while SVM classifies the data using these clusters. The accuracy of their models was measured as 97.38%. In the work presented in [27], Jabbar aimed to improve the accuracy of breast cancer diagnosis using a community learning approach. For this purpose, Bayesian Networks and Radial Basis Function are used in the proposed method. According to the findings, the proposed model reaches 97% classification accuracy. Similarly, Abdel-Zaher and Eldeib in [28] used Deep Belief Networks to diagnose breast cancer. According to the obtained results, their proposed system achieves an accuracy of 99.68%. In addition, Kamel et al. in [29] aimed to classify breast cancer data with the Gaussian Naive Bayes algorithm and achieved 98% accuracy.

In addition to these studies, artificial neural networks and deep learning-based models have been proposed for breast cancer diagnosis and classification, especially in recent years. In one of them, Alickovic and Subasi in [30] aimed to classify a breast cancer dataset using Normalized Neural Networks. According to the experimental results, it was found that their models achieved 99.27% classification accuracy. Similarly, Singh et al. [31] used a Feature Importance Score-Based Functional Link Artificial Neural Networks to classify the same dataset. The proposed model achieved classification with 99.41% accuracy. In addition, in the work given in [32], Kaur proposed a Dense Convolutional Neural Network-based framework for the same aim. His model's performance was also successful, similar to that of other ANN-based models. Its classification accuracy was 98.2%.

Along with the machine learning algorithms shared above, the kNN classification algorithm is also widely used for breast cancer diagnosis in various hybrid structures. In [33], one of these studies, Pawlovsky and Matsuhashi used Genetic Algorithm (GA) for component selection to improve the accuracy of kNN. For this purpose, they tried to make the GA select the best choromosomes. According to the experimental results, their proposed model achieved better results than the standard kNN on UCI's breast cancer dataset. While the standard kNN classifies the data with an accuracy of 76%, the proposed model classifies the data with an accuracy of 79%. In another study, Rajaguru and Chakravarthy [34] performed feature selection on breast cancer data using Principal Component Analysis and then classified the dataset using kNN and Decision Networks to compare the results. According to the results, kNN classified the dataset with 95.61% accuracy, while the Decision Tree classified it with 91.23%. In [35], Admassu performed hyperparameter optimization to determine the most accurate value of k for the kNN algorithm. The most appropriate k values for the breast cancer dataset were determined as 8 and 39, according to the findings. It was observed that the classification performance was 94.35% for the mentioned k values. Besides, in [36], Henderi et al. studied the effect of normalization on classification performance. They normalized the breast cancer dataset with Min-Max and Z-Score Normalization methods and classified it with kNN. The classification accuracy of their model was 98%. In [37], another study in this field, Tounsi et al. examined the effect of feature selection methods on the classification accuracy of breast cancer dataset. For this purpose, they classified the data using SVM and kNN after feature selection. The findings determined that kNN can classify with 96.83% when Ant Colony Optimization is used as the feature selection method. Another study aiming to classify breast cancer dataset with kNN and using feature selection was proposed by Priyadarshini et al. in [38]. The authors applied various feature selection methods to various datasets, including a breast cancer dataset, and then classified them with kNN. The results show 99.51% accuracy can be achieved when the kNN classification algorithm is used with the Equilibrium Optimizer.

As can be seen from the studies we have discussed in this section, clustering algorithms are underutilized for breast cancer diagnosis. This is one of the most critical drawbacks of existing studies. Specifically, density-based clustering algorithms like MCMST, which is very successful in detecting outliers, are anticipated to contribute to classification success significantly. To address this shortcoming in this area, this study is also crucial.

3. PRELIMINARIES

3.1. kNN (k-Nearest neighbors)

kNN is a simple and easy-to-understand algorithm with few parameters (k parameters representing only the k nearest neighbors) and high classification ability. As can be seen in Figure 1, kNN decides which class to assign the data to by looking at its k nearest neighbors. The dominant class among these k neighbors is determined as the class to which the relevant data will be assigned.

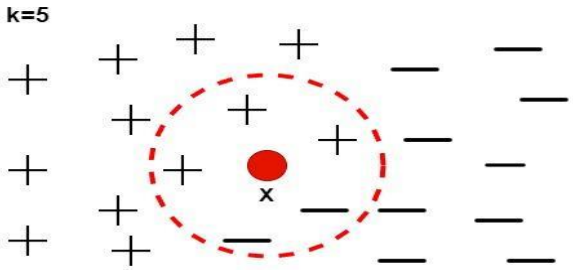

Figure 1. A kNN exmple (k=5).

3.2. Clustering Algorithms

Clustering algorithms are unsupervised learning algorithms that define clusters based on similarities and dissimilarities and do not require class labels. They are machine learning methods that can provide excellent results, especially when class labels are missing, incomplete, or inconsistent. For this purpose, they usually use the distances among the data as the similarity criteria. Euclidean distance, Mahalanobis, City Block, and Manhattan are commonly used distance calculation methods. K-means [17], DBSCAN [18], HDBSCAN [19], OPTICS [20], FuzzyCMeans [21], and MCMST [22] are some examples of these kinds of algorithms.

3.3. Outlier Detection

As shown in Figure 2, outliers refer to data that exhibit characteristics outside the normal. Outlier detection methods, also called anomaly detection, are proposed to detect such abnormal data. Various methods have been proposed to detect outliers, including IForest, MAD, IQR, and LOF. The main goal of such methods is to detect outliers through various mathematical and/or statistical calculations. Apart from such methods, clustering algorithms are also used to detect outliers. The main idea behind using clustering algorithms for detecting outliers is to detect data that fall outside

clusters by defining clusters. In particular, density-based clustering algorithms such as DBSCAN achieve very successful results in this regard.

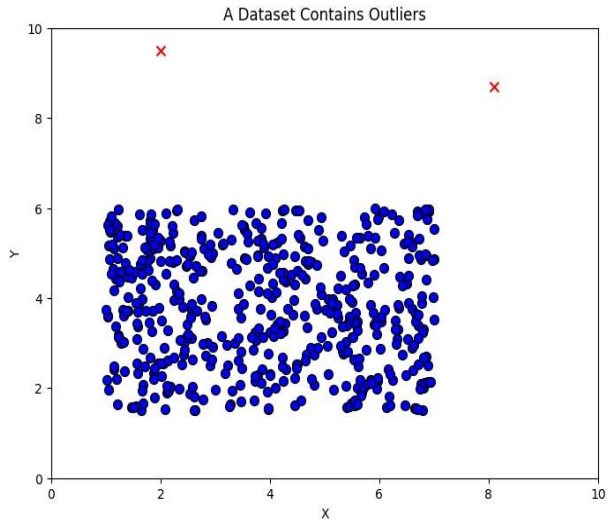

Figure 2. An example that shows outliers.

3.4. MCMST Algorithm

The MCMST algorithm is a density-based clustering algorithm that achieves high clustering success by using a KD-Tree data structure to define micro-clusters and then applying the Minimum Spanning Tree to these micro-clusters to identify the macro clusters [22]. As can be seen in Figure 3, since the MCMST algorithm is a density-based clustering algorithm, it is an algorithm that can both define clusters in arbitrary-shapes and detect outliers with high accuracy.

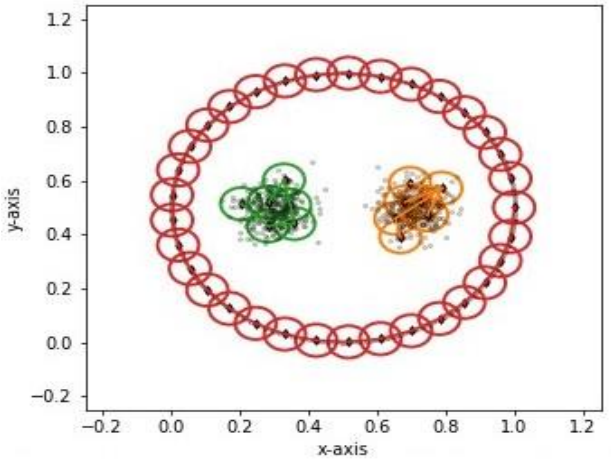

Figure 3. Clustering example with MCMST [22].

4. PROPOSED MODELS: CLUSTERING-BASED OUTLIER DETECTION AND KNN CLASSIFIER MODELS FOR BREAST CANCER DIAGNOSIS

This section provides detailed information about the proposed models. In this study, DBSCAN, HDBSCAN, OPTICS, FuzzyCMeans, and MCMST clustering algorithms were used as clustering algorithms. As seen in Figure 4, the proposed model first passes the dataset through an outlier filter using various clustering algorithms. Then, the data set, which has been cleaned from outliers, is subjected to classification with the kNN classifier. In the last stage, the obtained results are evaluated.

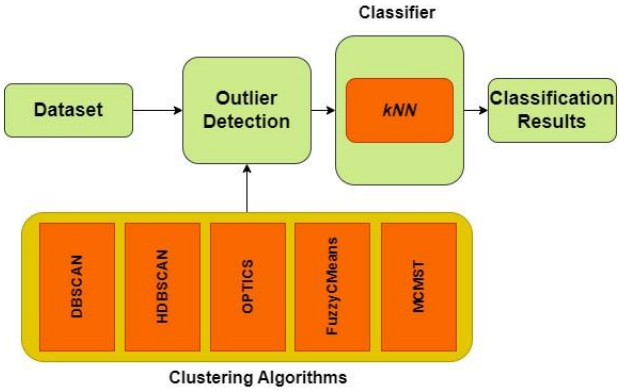

Figure 4. Proposed model.

4.1. Outlier Detection Using Clustering Algorithms

Clustering algorithms are methods that classify data into groups according to the similarities among them. In particular, density-based methods use various parameters for clustering. In short, an amount of data group with an enough density in a certain area that is above a certain threshold value are defined as clusters. Data groups that fall below this threshold are defined as outlier data. At this point of view, clustering algorithms are widely used in the data preprocessing stage to detect outliers. As illustrated in Figure 5, the process of eliminating the outliers is expected to have a positive effect on the clustering success. The first figure indicates the raw data with outliers, while the second figure illustrates the processed data. In this study, we use various clustering algorithms in the data preprocessing stage of classification algorithms to eliminate outliers and examine the possible impact of clustering algorithms on classification performance.

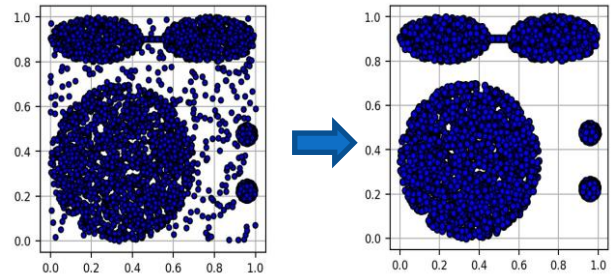

Figure 5. An example of outliers' elimination.

4.2. Classifier

In this study, we used the kNN as a classifier because it is easy to use, simple, and achieves successful results. Another important advantage of the kNN is that it uses only one parameter.

5. EXPERIMENTAL STUDY

In this section, we share detailed information about the experimental studies carried out to demonstrate the performances of the proposed models.

5.1. Used Dataset and Data Preprocessing

In this study, the Wisconsin Breast Cancer Dataset is used to measure the success of the proposed models. The dataset contains information about breast cancer diagnoses of 569 patients. Each record consists of 30 features. To make the parameter selection process easier, the data were normalized using Min-Max Normalization. Let x_{MinMax} be Min-Max Normalization of feature x that is the scaled value between [0, 1], x_{min} be the minimum value of x feature, and x_{max} be the maximum value of x; x_{MinMax} is calculated using Eq. (1).

$$x_{MinMax} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

5.2. Parameter Setting

Both clustering algorithms and kNN use various parameters. To understand which parameters achieve the best results, a random search method was performed for each algorithm. The range of parameters used is provided in Table 1. In addition, k-fold cross-validation was used to ensure that the results of the models were properly tested. Here, k is set to 5, and the dataset is divided into 75% training and 25% test data.

Table 1. Range of parameters.

Algorithm	Parameter range
kNN	k=[1, 20]
kNN + DBSCAN	k=[1, 20], eps=[0.01, 1], min_samples=[1, 30]
kNN + HDBSCAN	k=[1, 20], min_cluster_size=[1, 20], min_samples=[1, 30]
kNN + OPTICS	k=[1, 20], eps=[0.01, 1],, min_samples=[1, 30]
kNN + FuzzyCMeans	k=[1, 20], c=[1, 30], m=[1, 30]
kNN + MCMST	k=[1, 20], N=[1, 30], r=[0.01, 1],, n_micro=[1, 30],

5.3. Metrics to Measure the Classification Performance

Since a classification model was used in this study, the Accuracy, Precision, Recall, and F1-Score, commonly used metrics, were used to measure classification success. These metrics are calculated from the confusion matrix given in Table 2. Let True-Positive be TP, True Negative be TN, False Positive be FP, and False Negative be FN; each metric is calculated by Eq. (2), (3), (4), and (5), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

Table 2. Confusion matrix.

		Actual	
Predicted		TP	FP
		FN	TN

6. EXPERIMENTAL RESULTS AND DISCUSSION

In order to determine which clustering algorithm is more successful in detecting outliers, each clustering algorithm was run on the WBCD dataset with parameters randomly selected from the range given in Table 1. Then the results were analyzed by classifying with kNN. Each algorithm was run 100 times and the parameters given in Table 3 that provided the highest classification accuracy were determined. Each clustering algorithm detected a different number of outliers with these parameters, as shown in Table 4. Likewise, the classification successes shown in Table 5 were obtained when the models were tested with these parameters.

Table 3. The best parameters for each algorithm.

Algorithm	Parameters
kNN	k=12
kNN + DBSCAN	k=5, eps=0.54, min_samples=5
kNN + HDBSCAN	k=3, min_cluster_size=3, min_samples=2
kNN + OPTICS	k=16, eps=0.32, min_samples=13
kNN + FuzzyCMeans	k=8, c=13, m=2
kNN + MCMST	k=4, N=4, r=0.58, n_micro=24,

When the results are analyzed, it is seen that the MCMST + kNN model gives the highest classification performance. If the dataset was subjected to classification with kNN without using any outlier detection method, kNN classified the dataset with an accuracy of 0.9719. However, an accuracy rate of 0.9834 was achieved when the outliers were detected and deleted with the MCMST clustering algorithm on the dataset and then classified with kNN. If the effects of other clustering algorithms on the classification performance of kNN are analyzed, we can see that DBSCAN and FuzzyCMeans algorithms slightly increase the classification success, although not as much as MCMST. However, it is seen that HDBSCAN and OPTICS algorithms have a negative impact on classification performance.

Table 4. Number of detected outliers for the models.

Algorithm	# of Detected Outliers	Outliers Ratio (%)
kNN	-	-
kNN + DBSCAN	48	8.44
kNN + HDBSCAN	42	7.38
kNN + OPTICS	41	7.21
kNN + FuzzyCMeans	24	4.22
kNN + MCMST	7	1.23

Table 5. Comparison of classification performance of models.

	Accuracy	Precision	Recall	F1-Score
kNN	0.9719	0.9634	0.9771	0.9694
kNN + DBSCAN	0.9769	0.9702	0.9789	0.9742
kNN + HDBSCAN	0.9696	0.9596	0.9715	0.9651
kNN + OPTICS	0.9646	0.9495	0.9703	0.9584
kNN + FuzzyCMeans	0.9755	0.9677	0.9759	0.9714
kNN + MCMST	**0.9834**	**0.9796**	**0.9846**	**0.9817**

When kNN was applied to the dataset without performing any preprocessing related to outlier detection, the confusion matrices shown in Figure 6 were obtained. On the other hand, when the MCMST clustering algorithm, which had the highest classification accuracy, was used with kNN, the confusion matrices given in Figure 7 were obtained.

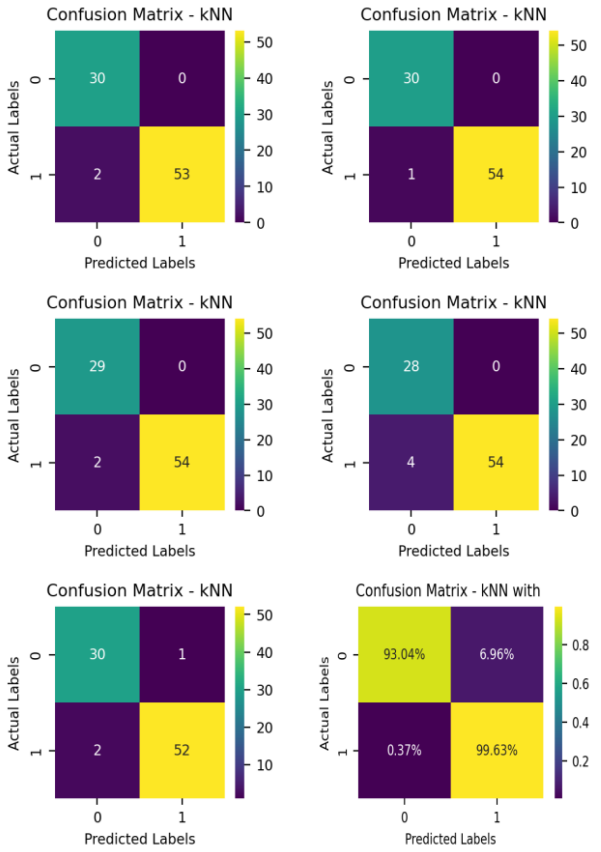

Figure 6. Confusion matrices and their average obtained from k-fold cross-validation for kNN (k=5).

Another point that we analyzed in the proposed study is the run-time complexity of models. As shown in Figure 8, the fastest model among others was the one in which the kNN was run alone. However, it should be noted that no outlier detection is performed in this model. In contrast, the model with the highest run-time was kNN + OPTICS. When the kNN + MCMST model, which achieves the highest classification success, is examined, it is seen that although it is slower than the other 4 models, it is considerably faster than the kNN + OPTICS model.

After determining that the model with the highest classification success was kNN + MCMST, we examined the effect of kNN's single parameter k on the classification success. We chose k from [1, 50] interval and ran the model for each value of k. The obtained results are illustrated in Figure 9. When the results are analyzed, it can be said that the k value affects model success, but it does not have a great effect. However, the value k = 4 gives the highest classification success.

The MCMST algorithm, as mentioned in the related study, is a very successful algorithm for detecting outliers. This is because this algorithm identifies clusters with a micro-cluster-based density approach. This approach makes the detection of outliers more efficient. Therefore, the MCMST + kNN model is expected to give better results than other models. The results obtained also support this expectation.

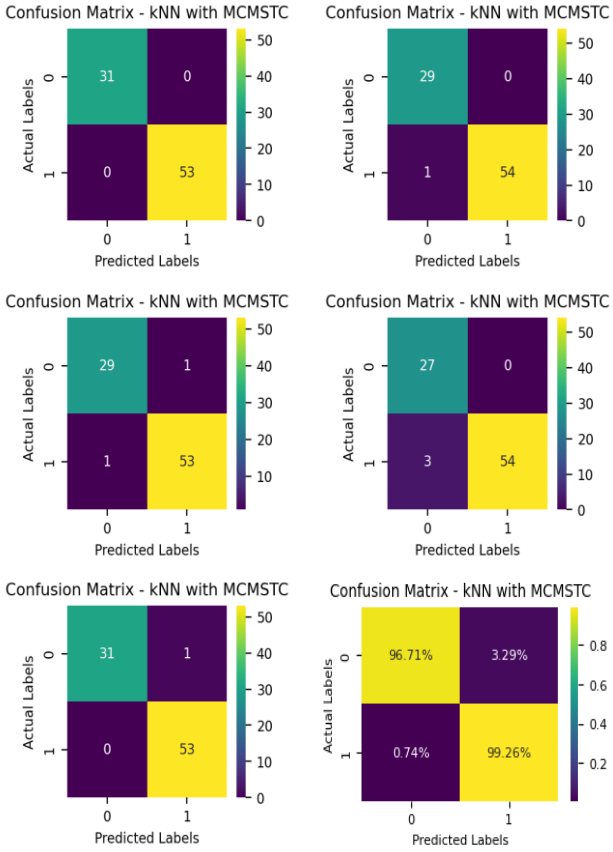

Figure 7. Confusion matrices and their average obtained from k-fold cross-validation for kNN + MCMST (k=5).

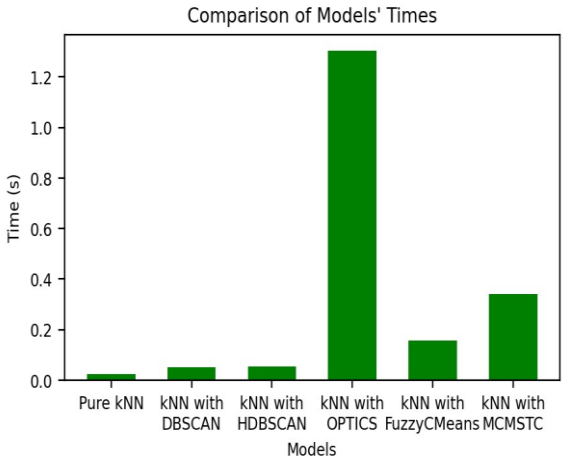

Figure 8. Comparison of the run-time complexity of the models.

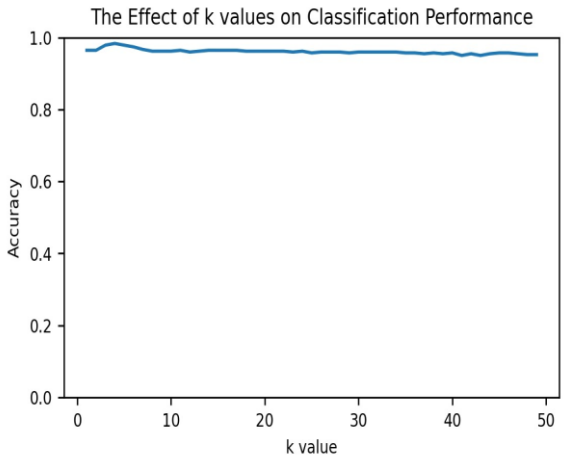

Figure 9. The effect of k values on the kNN + MCMST model classification.

7. CONCLUSION

In this study, the effect of using clustering algorithms in the data preprocessing stage to eliminate outliers on the classification accuracy of the WBCD dataset was investigated. For this purpose, the mentioned dataset was first processed with DBSCAN, HDBSCAN, OPTICS, FuazzyCMeans, and MCMST clustering algorithms to identify and eliminate outliers and then these data were separately classified with kNN that is a simple and effective algorithm.

According to the results, kNN + MCMST was the model with the highest classification performance. The highest classification accuracy was achieved by the kNN + MCMST model, which had an accuracy of 0.9834. In contrast, the accuracy of the kNN method without any data preprocessing was 0.9719. Although the run-time of the model is a little high, it is at an acceptable level. However, three predefined parameters for MCMST are difficult for regular users to determine. This is the most crucial limitation of the model. Given the success of the MCMST clustering algorithm in enhancing classification accuracy, future studies could investigate its integration with deep learning-based models. We can achieve even higher classification accuracy and robustness by leveraging the strengths of both clustering algorithms and deep learning architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). This hybrid approach could offer novel insights into the complex patterns underlying breast cancer diagnosis and contribute to the development of more effective diagnostic tools. The models to be developed are planned to be tested in different clinical data sets and in different fields. This will be important in testing the proposed model's effectiveness in different fields.

REFERENCES

[1] Sağlık, A. Rakamlarla Meme Kanseri. 2023 [cited 2023 12.09.2023]; Available from: https://www.anadolusaglik.org/blog/rakamlarla-meme-kanseri.

[2] Şenol, A., Canbay, Y. and Kaya, M., Trends in Outbreak Detection in Early Stage by Using Machine Learning Approaches. Bilişim Teknolojileri Dergisi. 14(4): p. 355-366.

[3] Khaire, U.M. and R. Dhanalakshmi, Stability of feature selection algorithm: A review. Journal of King Saud University-Computer Information Sciences, 2022. 34(4): p. 1060-1073.

[4] Zhou, H., X. Wang, and R. Zhu, Feature selection based on mutual information with correlation coefficient. Applied Intelligence, 2022: p. 1-18.

[5] Heidari, A., et al., Machine learning applications for COVID-19 outbreak management. Neural Computing Applications, 2022. 34(18): p. 15313-15348.

[6] Deiana, A.M., et al., Applications and techniques for fast machine learning in science. 2022. 5: p. 787421.

[7] Russell, S.J., Artificial intelligence a modern approach. 2010: Pearson Education, Inc.

[8] Manevitz, L.M. and M. Yousef, One-class SVMs for document classification. Journal of machine Learning research, 2001. 2(Dec): p. 139-154.

[9] Ali, N., D. Neagu, and P. Trundle, Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Applied Sciences, 2019. 1: p. 1-15.

[10] Fürnkranz, J., Decision Tree, in Encyclopedia of Machine Learning and Data Mining, C. Sammut and G.I. Webb, Editors. 2017, Springer US: Boston, MA. p. 330-335.

[11] Jain, A.K., J. Mao, and K.M. Mohiuddin, Artificial neural networks: A tutorial. J Computer, 1996. 29(3): p. 31-44.

[12] Liu, F.T., K.M. Ting, and Z.-H. Zhou, Isolation-Based Anomaly Detection. ACM Trans. Knowl. Discov. Data, 2012. 6(1): p. Article 3.

[13] Breunig, M.M., et al., LOF: identifying density-based local outliers. SIGMOD Rec., 2000. 29(2): p. 93–104.

[14] Schölkopf, B., et al., Estimating the support of a high-dimensional distribution. Neural Computation, 2001. 13(7): p. 1443-1471.

[15] Rousseeuw, P.J. and C. Croux, Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association, 1993. 88(424): p. 1273-1283.

[16] Ahmad, S., et al., On efficient monitoring of process dispersion using interquartile range. Open journal of applied sciences, 2012. 2(04): p. 39-43.

[17] Hartigan, J.A. and M.A. Wong, A k-means clustering algorithm. JSTOR: Applied Statistics, 1979. 28(1): p. 100--108.

[18] Ester, M., et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996, AAAI Press: Portland, Oregon. p. 226-231.

[19] Campello, R.J.G.B., D. Moulavi, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. in Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2013.

[20] Ankerst, M., et al., OPTICS: ordering points to identify the clustering structure. SIGMOD Rec., 1999. 28(2): p. 49–60.
[21] Bezdek, J.C., R. Ehrlich, and W. Full, FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984. 10(2): p. 191-203.
[22] Şenol, A., MCMSTClustering: defining non-spherical clusters by using minimum spanning tree over KD-tree-based micro-clusters. Neural Computing and Applications, 2023. 35(18): p. 13239-13259.
[23] Chen, H.-L., et al., A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst. Appl., 2011. 38(7): p. 9014–9022.
[24] Marcano-Cedeño, A., J. Quintanilla, and D. Andina, WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications, 2011. 38: p. 9573-9579.
[25] Seera, M. and C.P. Lim, A hybrid intelligent system for medical data classification. Expert Systems with Applications, 2014. 41(5): p. 2239-2249.
[26] Zheng, B., S.W. Yoon, and S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 2014. 41(4, Part 1): p. 1476-1482.
[27] Jabbar, M.A., Breast Cancer Data Classification Using Ensemble Machine Learning. Engineering and Applied Science Research, 2021. 48(1): p. 65-72.
[28] Abdel-Zaher, A.M. and A.M. Eldeib, Breast cancer classification using deep belief networks. Expert Systems with Applications, 2016. 46: p. 139-144.
[29] Kamel, H., D. Abdulah, and J.M. Al-Tuwaijari. Cancer Classification Using Gaussian Naive Bayes Algorithm. in 2019 International Engineering Conference (IEC). 2019.
[30] Alickovic, E. and A. Subasi. Normalized Neural Networks for Breast Cancer Classification. in CMBEBIH 2019. 2020. Cham: Springer International Publishing.
[31] Singh, S., et al., Feature Importance Score-Based Functional Link Artificial Neural Networks for Breast Cancer Classification. BioMed Research International, 2022. 2022: p. 2696916.
[32] Kaur, H., Dense Convolutional Neural Network Based Deep Learning Framework for the Diagnosis of Breast Cancer. Wireless Personal Communications, 2023.
[33] Pawlovsky, A.P. and H. Matsuhashi. The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis. in 2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE). 2017.
[34] Rajaguru, H. and S. Chakravarthy, Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. Asian Pacific journal of cancer prevention : APJCP, 2019. 20: p. 3777-3781.
[35] Admassu, T., An optimized K-Nearest Neighbor based breast cancer detection. Journal of Robotics and Control (JRC), 2021. 2.
[36] Henderi, H., Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. IJIIS: International Journal of Informatics and Information Systems, 2021. 4: p. 13-20.
[37] Tounsi, S., I.F. Kallel, and M. Kallel. Breast cancer diagnosis using feature selection techniques. in 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). 2022.
[38] Priyadarshini, J., et al. Analyzing Physics-Inspired Metaheuristic Algorithms in Feature Selection with K-Nearest-Neighbor. Applied Sciences, 2023. 13(2), 906.