Applied Mathematics and Computational Intelligence
Volume 12, No. 4, Nov 2023 [1-12]

Comparison of Classification Models for Breast Cancer Disease Using Multivariate Analysis and Data Mining Approaches

Nurul Ashyikin Ramli[1], Zalina Zahid[2*], Siti Aida Sheikh Hussin[3] and Noor Asiah Ramli[4]

[1,2,3,4]College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Shah Alam, 40450, Shah Alam, Selangor, Malaysia.

[*]Corresponding author: zalina@tmsk.uitm.edu.my

Received: 27 Oct 2022
Accepted: 8 Sept 2023

ABSTRACT

Compared to other cancer types, breast cancer is one of the main causes of death in women. Early cancer detection can significantly increase survival and quality of life. A variety of machine learning prediction algorithms with combination of feature selection approaches have shown to be useful in the detection of breast cancer disease. However, it was discovered that there are still problems with classification accuracy. An outlier-related factor was known to have potential effect on classification accuracy. In order to further improve the classification's accuracy, the K-means approach was used to detect outliers. The major goal of this study was to examine the classification performance of breast cancer disease when feature selection methods were used in combination with K-Means. For experimental purpose, the Coimbra dataset for breast cancer consisting of 116 instances and 10 attributes was used in this study. Multivariate techniques including Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), and Discriminant Analysis (DA) were applied to reduce data dimensions. Meanwhile, four data mining approaches consisting of Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) were compared for classification purpose. The performance measurement was then evaluated using accuracy, precision, specificity, and sensitivity criteria. The results revealed that five combinations approaches (PCA-DT, PCA-RF, KPCA-DT, KPCA-RF, DA-RF) performed better across all four criteria after combining with K-Means technique. Among five combined methods, KPCA with DT outperformed other combination methods with the highest value across precision (76.47 percent) and specificity (71.43 percent). This study suggests the incorporation of feature selection method together with outlier detection method has proved to be more efficient and beneficial for breast cancer classification.

Keywords: Breast Cancer, Principal Component Analysis, Kernel Principal Component Analysis, Random Forest, Support Vector Machine.

1 INTRODUCTION

Breast cancer is the most common cancer affecting women worldwide and is the second most serious of all cancer diseases [1]. Not only breast cancer interferes with people's daily lives, but it also places a heavy financial load on them due to expensive medical expenses [2]. Early breast cancer diagnosis and identification are crucial for successful therapy as it helps clinically tailor preventative and

treatment strategies, lower the disease's recurrence rate, improve patient prognoses, and lengthen patients' lives [3].

Modern machine learning classifiers, according to Ibrahim et al. [4], can enhance early breast cancer tumour identification. In recent years, a variety of machine learning prediction algorithms have shown useful in the detection of diseases, and more intelligent prediction results have proven useful in helping clinicians diagnose diseases more quickly and accurately. The disease of breast cancer has been studied so far using a wide range of data mining techniques, which are the sub-components of machine learning approaches. The most widely used data mining classification methods include decision trees, random forests, and support vector machines [1,5,6]. However, these data mining techniques which assess every characteristic of breast cancer data, ignores the impact of redundant data on the outcomes of experiments as well as the interrelationships between attribute components. To improve the data mining classification technique, other methods, such as feature selection method, have been used.

But even with the application of the feature selection method, the accuracy still is not satisfactory. It was known that an outlier-related factor could affect classification accuracy. In order to further improve accuracy for breast cancer patients, outliers were detected using the K-means approach. In other words, not every problem can be solved optimally by the best algorithm if there are other factors that can affect classification accuracy [7]. K-means clustering was therefore incorporated into this study in order to improve it even more. KPCA was also implemented into this analysis to account for the potential for non-linear relationships between the variables. This paper identified the best feature selection techniques combined with K-means and compared four data mining classification method that produced the best performances in classifying the breast cancer cases.

The disease classification of breast cancer has been studied so far using a wide range of data mining techniques, which are the sub-components of machine learning approaches. The most widely used data mining classification methods include decision trees, random forests, and support vector machines [1,5,6]. As some of the models were categorized as linear and the others as non-linear, it was revealed that their accuracy performance varies among one another [8]. The IBk, Bagging, Random Forest, Random Committee, and SimpleCART algorithms were the most successful algorithms, according to the findings, with above 90% detection accuracy.

These data mining techniques which assess every characteristic of breast cancer data, also ignore the impact of redundant data on the outcomes of experiments as well as the interrelationships between attribute components. To improve the data mining classification technique, other methods, such as feature selection method, have been used. There are three types of feature selection methods which are supervised, semi-supervised, and unsupervised. Unsupervised feature selection methods are generally considered to be a more unbiased approach that can perform very well in addition to being able to reduce the risk of data-overfitting compared to supervised and semi-supervised feature selection method [9]. Among the best feature selection methods, PCA and DA are recommended for high dimensional data which is applicable to breast cancer data. Iqbal et al. [10] applied PCA, discriminant analysis, and logistic regression to minimize the dimensions in the Wisconsin dataset for breast cancer. For all three of the employed feature selection methods, they discovered that the support vector machine performed better than other approaches. Additionally, they discovered that a support vector machine feature selection strategy using discriminant analysis had the maximum

accuracy. The feature selection technique is also employed with the kernel PCA extension method of PCA. Meanwhile, Mushtaq et al. [11] employed this technique to condense the dimension spaces in their research. They discovered that using kernel PCA improved accuracy performance when compared to results produced by earlier researchers.

Das and Mohanty [12] also chose features for their investigation using this method. They chose to utilise this technique because KPCA can denoise data well. Identifying outliers in data is a crucial step in data analysis and eliminating them from clusters can increase the clustering's accuracy [13]. The most often used approach for identifying outliers is the K-Mean distance-based method. A technique for grouping or dividing a pattern into several clusters so that related patterns are assigned to the same cluster is called K-means clustering [14]. The method can cluster data and identify outliers at the same time. The cluster centre computation does not take outliers into account. In a study by Barai and Dey [15] on the identification of outliers using K-means and hierarchical clustering, the researchers discovered that accuracy increased after the outliers were removed using the K-means approach.

2 METHODOLOGY

This section describes the research methodology and statistical analyses employed in the study.

2.1 Research Framework

According to Figure 1, which depicts the study's research flow, the suggested model's mechanism passes through five primary stages: data pre-processing, feature selection, clustering, classification, and performance evaluation. The data used in this work was taken from the Breast Cancer Coimbra dataset, which was added to the UCI Machine Learning repository in 2018. This data set contains 116 cases (45% healthy controls and 55% breast cancer patients), 10 clinical features (age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1), and one binary dependent variable (identifying the presence or absence of breast cancer).

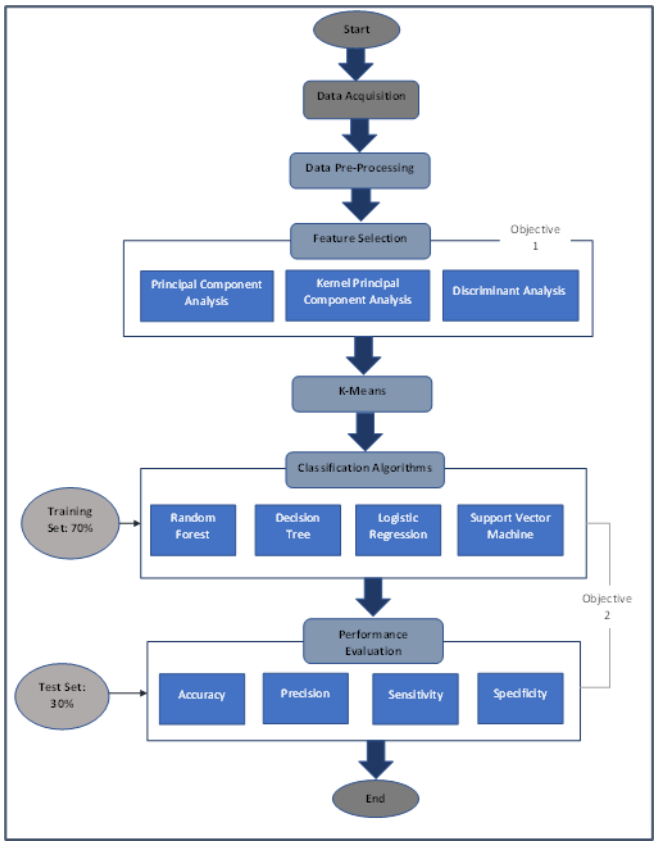

Figure 1: Research flow

To enhance the quality of the data and provide a clean dataset that could be utilised to develop the model, data pre-processing was carried out. We used several data processing approaches, including outlier detection, data cleaning, and data normalisation, to make our initial dataset more useful and usable for predicting breast cancer. Three different dimensionality selection techniques used and compared in this study. Principal Component Analysis, the first technique, is the most widely used dimensionality selection technique (PCA). The second method is Kernel Principal Component Analysis (KPCA) followed by Discriminant Analysis (DA). Kernel Principal Component Analysis (KPCA), followed by Discriminant Analysis (DA), is the following technique.

2.2 Feature Selection

2.1.1 Principal Component Analysis

The process of feature selection involves basically translating the initial feature space to a low-dimensional feature space through the relationship between traits in order to accomplish the goal of dimension selection. The correlation problem, which makes it difficult for the classification algorithm to identify correlations among the data, is then overcome by using PCA to modify the initial collection of features [16]. As an unsupervised learning dimensionality selection technique, PCA reduces the

data dimension by associating multidimensional data groupings. It might reduce the computation cost of the algorithm by reducing information loss, simplify the data structure, make the data set easier to use, completely without parameter restrictions, and create the data set [17].

2.1.2 Kernel Principal Component Analysis (KPCA)

The calculated covariance matrix C from the input data is transformed using PCA as its foundation. However, PCA is better suited for use in linear systems, thus this technique is ineffective for nonlinear data. With KPCA, which combines the linear PCA and the Kernel technique, nonlinear systems can perform better. The fundamental principle of kernel PCA is to perform a non-linear mapping $\Phi: \mathbb{R}^p \rightarrow F, y \mapsto Y$, as shown in Figure 2 below, to determine the principal component scores in higher dimensional space.

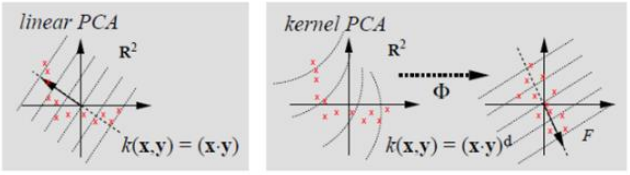

Figure 2: Illustration of Kernel PCA (Source: Ahsan et al., 2022)

Consider that the centred data are mapped to feature space *F*, $\Phi(x_1), \ldots, \Phi(x_n)$. The *n x n* feature space covariance matrix can be expressed as follows:

$$C^F = \frac{1}{n}\sum_{j=1}^{n} \Phi(x_j)\Phi(x_j)^T \quad (1)$$

Finding the eigenvalues of the eigenvector with eigenvalue $\lambda \geq 0$ that fulfils equations below.

$$C^F V = \lambda V \quad (2)$$

$$\lambda V = \frac{1}{n-1}\sum_{i=1}^{n} \phi(x_i)(\phi(x_i)^T V) \quad (3)$$

2.1.3 Discriminant Analysis (DA)

Fisher [24] developed the technique of discriminant analysis in 1936. Another name for it is Fisher Discriminant Analysis. The objective of DA is to combine the initial predictors to produce a new variable. In order to do this, the differences between the predefined groups with respect to the new variable are maximized [18]. It is assumed that the dataset is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ and that any sample x_i is an n-dimensional vector with $x_i = \{C_1, C_2, .., C_k\}$. $X_j (j = 1,2, \ldots, K)$ is a group of class *j* samples, and $\mu_j (j = 1,2, \ldots, K)$ is the mean vector of the *j* sample. We define N_j *(j = 1, 2, ..., K)* as the number of samples of class *j*. Define the covariance matrix for the class *j* samples as $\Sigma_j (j = 1,2, \ldots, K)$. The μ_j and Σ_j can be calculated using this equation:

$$\mu_j = \frac{1}{N_j} \sum_{x \epsilon X_j} x \tag{4}$$

$$\Sigma_j = \sum_{x \epsilon X_j} (x - \mu_j) \tag{5}$$

Discriminant analysis's main objective is to separate samples from different groups. In essence, it changes data into a different space that recognizes classes that can be referred to as "between classes (S_b)" and "within classes (S_w)" optimally. Equation below indicated S_b and S_w where μ_k is the class k mean and μ is the overall average. S_t=S_b+ S_w is the formula for the total covariance matrix. The main goal is to maximize between-class scatter, S_b, while minimizing within-class scatter, S_w. This involves separating different classes as much as is practicable.

$$S_b = \sum_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{6}$$

$$S_w = \sum_k \sum_{i \epsilon k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{7}$$

2.3 Outlier Detection

K-means is one of the most straightforward and efficient unsupervised classification methods. A popular partitioning-based clustering method called K-means seeks out a predetermined number of clusters that can be represented by their centroids. The distance between items is utilized as a measure of similarity in this conventional distance-based clustering technique, and the smaller the distance, the more similar the objects are [19].

2.4 Classification Algorithms

Building a classification model from a given data set that includes some attributes and labelled classes is the goal of supervised learning. Two essential parts that are used in supervised learning are the training data set and the testing data set. The prediction model is constructed using the training data set, which also contains attributes and cluster values. In this work, testing was done using Random Forest (FR), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). These models were chosen for their performance and popularity in literature.

i. Decision Tree (DT)

Decision Tree (DT) is a simple and straightforward classifier. Only Decision Trees offer the bit through feature to access detailed patient details. Decision trees construct classification or regression models using a tree-like structure that makes them easy to use and debug. Both category and numerical data can be processed using decision trees. Finding the information gain of the attributes and removing them allows the algorithm to break the branches into threes [20].

ii. Random Forest (RF)

Decision trees are used to combine tree predictors in random forests (RF), where each tree is dependent on values from a random vector that was sampled consistently and randomly over the entire forest [20]. The strength of each individual tree in the forest and the correlation between them determine the generalization error of a forest of tree classifiers. They are more noise-resistant and sturdy. It is a supervised classification technique used for prediction, and it is regarded as the best since it uses a lot more trees than decision trees, which results in higher accuracy.

iii. Support Vector Machine (SVM)

Support vector machines are learning tools that use a high-dimensional feature space with a hypothesis linear function space. They are taught using a learning method based on optimization theory that was developed from statistical learning theory [21]. SVM can only be used to data sets that have precisely two groups to categorize. By selecting the optimum hyperplane to divide all data points into one of two groups, it categorizes the data.

iv. Logistic Regression (LR)

The biological sciences are among the numerous domains that have made extensive use of the logistic regression model [19]. When categorizing data objects into groups is the goal, the logistic regression approach is utilized. In logistic regression, the target variable is typically binary, which means that it only contains data that can be classified as 1 or 0. In our case, this refers to a patient's breast cancer stage, either healthy control or patient, and is determined by several factors.

2.9 Performance Measures

The efficiency of the suggested approach is assessed by considering the actual and projected categorization. In order to determine the system's accuracy, the confusion matrix for the chosen classifier is employed [22]. The best classification methods that can be used to predict breast cancer were compared and identified in the study's final phase using performance measures like accuracy, precision, specificity, and sensitivity. Due to the popularity of this performance metric among prior researchers, it was chosen [1,6,21].

3 RESULTS

An excessive value known as an outlier may have an impact on the analysis. Given that k-means clustering can deal with outlier, we used it to solve the outlier problem. The dataset has five extreme values that have been recognized and need to be dealt with. In order to determine the similar and dissimilar group, K-means clustering was applied. 111 observations are in a similar group, whereas only 5 are in the dissimilar group. The dissimilar group was later labelled as an outlier and

disregarded. The data mining methods were then examined using the cleansed dataset. As a result, 111 observations were the final observations used in the analysis. The similar group (cluster 0) and dissimilar group (cluster 1) were displayed in Figure 3 below.

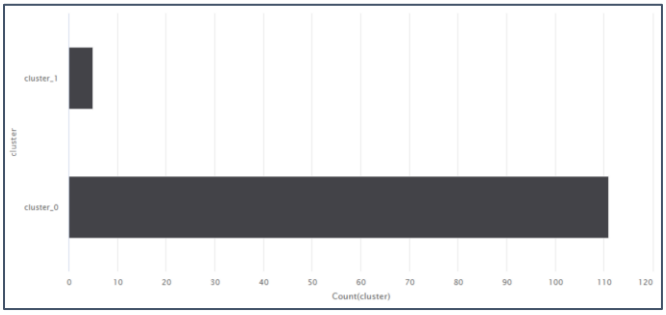

Figure 3: Similar and Dissimilar Group

After performing all feature selection approaches, classification methods, and K-means, the overall results were compared to determine which feature selection and classification method performed the best overall. The comparison of all classification performance without and with K-Means clustering applied shown in Table 1. Across all twelve combination of feature selection methods and classification models, it is identified that there are five combinations (PCA-DT, PCA-RF, KPCA-DT, KPCA-RF, DA-RF) that were found to be better across all four criteria after combining with K-Means cluster.

Table 1: Overall Performance Measurement Without and with K-Means Application

COMBINATION METHODS	**Without K-Means**				**With K-Means**			
	Accuracy	**Precision**	**Specificity**	**Sensitivity**	**Accuracy**	**Precision**	**Specificity**	**Sensitivity**
PCA + DT	62.86	63.64	50	73.68	**76.53**	**73.68**	**69.75**	**78.77**
PCA + RF	61.76	63.16	56.25	66.67	**71.43**	**72.66**	**68.73**	**73.68**
PCA + SVM	62.86	71.43	75	52.63	**70.59**	66.67	50	**88.89**
PCA + LR	62.86	66.67	62.50	63.16	**73.53**	**71.43**	62.50	**83.33**
KPCA + DT	57.14	60	50	63.16	**74.19**	**76.47**	**71.43**	**76.47**
KPCA + RF	60	61.90	50	68.42	**70.97**	**70**	**57.14**	**82.35**
KPCA + SVM	54.29	58.82	56.25	52.63	**71.14**	**68.18**	50	**88.24**
KPCA + LR	62.86	63.64	50	73.68	**64.52**	**68.75**	**64.29**	68.75

DA + DT	57.14	62.50	62.50	52.63	**70.59**	**70**	62.50	**77.78**
DA + RF	57.14	59.09	43.75	68.42	**70.59**	**68.18**	**56.25**	**83.33**
DA + SVM	62.86	62.50	43.75	78.95	**67.65**	**62.96**	37.50	**94.44**
DA + LR	65.71	65.22	50	78.95	64.71	63.64	50	77.78

Furthermore, another five combinations which are PCA-LR, KPCA-SVM, KPCA-LR, KPCA-DT, and DA-SVM were found to be better across three criteria after combining with K-Means clustering. While there is one combination that is better across two criteria after combining with K-means clustering which is PCA-SVM. Lastly, there is one combination had underperformed across all the criteria performances after combining with K-Means cluster which are DA-LR. Across eleven combinations, it can be shown that they have performed based on at least two criteria of performance. As for DA-LR, it is indicated that the application of k-means clustering is not improving since a LR is a linear classification method. It is consistent with the findings Ahsan et al. [23] that linear regression commonly known to has poor performance compared to the non-linear classification methods. Thus, this suggests the application of K-Means clustering has improved the classification process. This finding proved the results obtained by Barai and Dey [15] when they discovered that accuracy increased after the outliers were removed using the K-means approach.

Among five combinations which performed after the application of k-means methods, KPCA-DT scored the highest value of across two criteria which are precision and specificity as compared to the other four methods. Furthermore, KPCA-DT scored a slightly lower value of accuracy as compared to PCA-DT which has the highest value of accuracy. Therefore, it can be considered that KPCA and DT is the best method. The combination methods with the highest performance shown in Table 2.

Table 2: The Highest Performance Measurements for Five Combination Methods

COMBINATION METHODS	Without K-Means				With K-Means			
	Accuracy	Precision	Specificity	Sensitivity	Accuracy	Precision	Specificity	Sensitivity
PCA + DT	62.86	63.64	50	73.68	**76.53**	73.68	69.75	78.77
PCA + RF	61.76	63.16	56.25	66.67	71.43	72.66	68.73	73.68
KPCA + DT	57.14	60	50	63.16	74.19	**76.47**	**71.43**	76.47
KPCA + RF	60	61.90	50	68.42	70.97	70	57.14	82.35
DA + RF	57.14	59.09	43.75	68.42	70.59	68.18	56.25	**83.33**

From the above tables, it can be seen that the performance of the feature selection technique using kernel PCA and decision tree outperforms other combinations of techniques since it has the greatest value for precision and specificity, while ranking second for accuracy performance. PCA-DT, which yields the best accuracy as well as the second-highest values for precision and specificity, is the second approach that performs well. Finally, when it comes to sensitivity, DA-SVM performed at 94.44 percent, which is the highest among the other methods. However, discriminant analysis underperformed in terms of accuracy, precision, and specificity when compared to PCA and KPCA.

4 CONCLUSION

This study adds to the body of knowledge by recommending classification methods for breast cancer prediction. In order to ensure that the classification algorithms yield accurate results, we also determined the optimal feature selection method that may be utilized to minimize the dimensions. The study is useful for increasing the accuracy of classifiers by analyzing data using the K-means approach. This work will help fill the gap in the literature because there are not many studies employing K-means to analyze outlier identification in breast cancer prediction studies. To achieve better findings, the next researcher can apply other high level of unsupervised feature selection techniques to other high dimensional datasets such as Minimum Redundancy Maximum Relevance. Additionally, to provide more precise and effective findings for future work, the discriminant analysis and logistic regression may be integrated with other types of multivariate normality tests to improve its classification accuracy.

ACKNOWLEDGEMENTS

The authors would like to thank a great appreciation to Universiti Teknologi MARA (UiTM) Shah Alam for its support upon the completion of this research. We would also like to express our appreciation to the ISTEC-CoSQA 2023 for the opportunity to present our research findings.

REFERENCES

[1] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", *SN Computer Science*, vol. 1, no. 5, 1–14, 2020. Retrieved from https://doi.org/10.1007/s42979-020-00305-w

[2] K. Bian, M. Zhou, F. Hu, and W. Lai, "RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction. Frontiers in Genetics", 11(September), 2020. Retrieved from https://doi.org/10.3389/fgene.2020.566057

[3] R. K. Charaghvandi, B. van Asselen, M. E. P. Philippens, H. M. Verkooijen, C. H. van Gils,P. J. van Diest, R. M. Pijnappel, M. G. G. Hobbelink, A. J. Witkamp, T. van Dalen, E. van der Wall, T. C. van Heijst, R. Koelemij, M. van Vulpen, and H. J. G. D. van den Bongard, " Redefining radiotherapy for early-stage breast cancer with single dose ablative treatment: A study protocol", *BMC Cancer*, vol. 17, no. 1, 1–9, 2017. Retrieved from https://doi.org/10.1186/S12885-017-3144-5/FIGURES/4

[4] S. Ibrahim, S. Nazir, and S. A. Velastin, "Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis", *Journal of Imaging*, vol. 7, no. 11, 2021. Retrieved from https://doi.org/10.3390/jimaging7110225

[5] J. S. Ravi Shankar, S. Nithish, M. Nithish Babu, R. Karthik, and A. Shahid Afridi, "Breast Cancer Prediction using Decision Tree", *Journal of Physics: Conference Series*, vol 1916, no. 1, 2021.

Retrieved from https://doi.org/10.1088/1742-6596/1916/1/012069

[6] S. Sohrabi, and A. Atashi, "Prediction Breast Cancer Risk: Performance Analysis Data Mining Techniques", *Frontiers in Health Informatics*, vol. 10, no. 1, 83, 2021. Retrieved from https://doi.org/10.30699/fhi.v10i1.296

[7] B. Çiğşar, and D. Ünal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk. Scientific Programming, 2019". Retrieved from https://doi.org/10.1155/2019/8706505

[8] L. D. Arancibia, P. Sánchez-González, E. J. Gómez, M. E. Hernando, and I. Oropesa, "Linear vs Nonlinear Classification of Social Joint Attention in Autism Using VR P300-Based Brain Computer Interfaces", MEDICON 2019, *IFMBE Proceedings 76*, vol 1, 1869–1874, 2020. Retrieved from https://doi.org/10.1007/978-3-030-31635-8

[9] S. Solorio-fernández, and A. Carrasco-ochoa, "Unsupervised Feature Selection Method for Mixed Data", Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE), Technical Report No. CCC-19-005, 2019.

[10] T. Iqbal, A. Farooq, N. Sarwar, M. Ashraf, and A. Irshad, "Prediction of Breast Cancer Using Machine Learning Techniques", *BioScientific Review (BSR),* vol. 4, no. 1, 2022. Retrieved from https://doi.org/https://doi.org/10.32350/bsr

[11] Z. Mushtaq, M. F. Qureshi, M. J. Abbass, and S. M. Q. Al-fakih, "Effective kernel-principal component analysis-based approach for Wisconsin breast cancer diagnosis", vol. 59, no. 2, 1–4, 2023.

[12] A. Das, and M. N. Mohanty, "Design of ensemble recurrent model with stacked fuzzy ARTMAP for breast cancer detection", *Applied Computing and Informatics*, 2022. Retrieved from https://doi.org/10.1108/ACI-03-2022-0075

[13] G. Gan, and M. K. P. Ng, "k-means clustering with outlier removal", Pattern Recognition Letters, 90, 8–14, 2017. Retrieved from https://doi.org/10.1016/J.PATREC.2017.03.008

[14] S. H. Abdulla, A. M., and H. Veisi, "Breast cancer segmentation using K-means clustering and optimized region-growing technique", *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, 158–167, 2022. Retrieved from https://doi.org/10.11591/eei.v11i1.3458

[15] D. A. Barai and L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering", *World Journal of Computer Application and Technology*, vol. 5, no. 2, 24–29, 2017. Retrieved from https://doi.org/10.13189/wjcat.2017.050202

[16] C. Zhu, C. U. Idemudia, and W. Feng, "Improved Logistic Regression Model for diabetes prediction by integrating PCA and K-means techniques", *Informatics in Medicine Unlocked*, 17(March), 100179, 2019.

[17] A. S. Hess and J. R. Hess, "Principal component analysis", *Transfusion*, vol. 58, no. 7, 1580–1582, 2018. Retrieved from https://doi.org/10.1111/TRF.14639

[18] M. Keleş, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study", *Tehnicki Vjesnik*, vol. 26, no. 1, 149–155, 2019. Retrieved from https://doi.org/10.17559/TV-20180417102943

[19] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining | Elsevier Enhanced Reader", Informatics in Medicine Unlocked, 10, 100–107, 2018. Retrieved from https://reader.elsevier.com/reader/sd/pii/S2352914817301405?token=AD33AB33D600300375407EE3E7125B78A19328D8EF94BA4E7E89F9317131FA3862AE24BBDD8E6194C524F18F6B8E1916&originRegion=eu-west-1&originCreation=20220122003256

[20] H. B. F. David and S. A. Belcy, "Heart Disease Prediction Using Data Mining Techniques", *ICTACT Journal On Soft Computing*, vol. 9, no. 1, 1817–1823, 2018. Retrieved from https://doi.org/10.21917/ijsc.2018.0253

[21] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, and Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction.", *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 9, 192–201, 2018. Retrieved from https://doi.org/10.24843/lkjiti.2018.v09.i03.p08

[22] M. Kumari and V. Singh, "Breast Cancer Prediction System", Procedia Computer Science, 132, 371–376, 2018. Retrived from https://doi.org/10.1016/j.procs.2018.05.197

[23] M. Ahsan, M. Mashuri, H. Khusna, and Wibawati, "Kernel principal component analysis (PCA) control chart for monitoring mixed non-linear variable and attribute quality characteristics", *Heliyon*, vol. 8, no. 6, e09590, 2022. Retrieved from https://doi.org/10.1016/j.heliyon.2022.e09590

[24] E. M. Fisher, "Linear Discriminant Analysis. Statistics & Discrete Methods of Data Sciences", 392, 1-5, 1936.