Volume 13, Issue 3, Page 109-116, 2024 https://doi.org/10.46810/tdfd.1460871 Research Article

BİNGÖL UNIVERSITY

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

Evaluation of Missing Data Imputation Methods and PCA Techniques for Machine Learning Models in Breast Cancer Diagnosis Using WBCD

Yavuz Bahadır KOCA[1*], Elif AKTEPE[2]

[1] Afyon Kocatepe University, Engineering Faculty, Electrical Engineering Department, Afyonkarahisar, Türkiye
[2] Afyon Kocatepe University, Afyon Vocational School, Electronics and Automation Department, Afyonkarahisar, Türkiye
Yavuz Bahadır KOCA ORCID No: 0000-0002-0317-1417
Elif AKTEPE ORCID No: 0000-0002-2375-2040

**Corresponding author: ybkoca@aku.edu.tr*

(Received: 28.03.2024, Accepted: 25.08.2024, Online Publication: 26.09.2024)

Keywords
Breast cancer, Machine learning, Missing data management, PCA, Biomedical

Abstract: Cancer is one of the leading causes of human death, and breast cancer deaths are widespread among women. Early diagnosis of breast cancer is considered one of the main ways to reduce these deaths. Expert systems, artificial intelligence (AI), and machine learning (ML) techniques aim to assist doctors in the medical field in early disease detection. One of the main purposes of these technologies is to diagnose life-threatening diseases such as breast cancer earlier and accurately. This study analyses the Wisconsin Breast Cancer Dataset (WBCD) and evaluates the effects of different missing data imputation methods with Principal Component Analysis (PCA). PCA-based data reduction techniques are used in supervised ML methods. This study emphasizes combining ML approaches with missing data management strategies for breast cancer diagnosis. The study dataset consists of 699 data. 16 out of the data were identified as missing data. These missing data were processed using different data imputation methods. It was seen that the median filling technique provided the best performance. After filling the missing data with the median, PCA-based data reduction techniques were used on the dataset. The performances of Decision Trees (DT), Linear Regression (LR), Logistic Regression (LogR), k Nearest Neighbours (k-NN), Polynomial Regression (PR), Random Forest (RF) and Support Vector Machines (SVM) models were investigated for classifying tumours. The effects of these techniques were examined on ML algorithms with various PCA component numbers. The best performance was observed in SVM and k-NN algorithms. The success rates were 97.14% and 98.57%, respectively.

WBCD Kullanılarak Meme Kanseri Tanısında Makine Öğrenme Modelleri İçin Eksik Veri Atama Yöntemlerinin ve PCA Tekniklerinin Değerlendirilmesi

Anahtar Kelimeler
Meme kanseri, Makine öğrenmesi, Eksik veri yönetimi, PCA, Biyomedikal

Öz: Kanser, insan ölümlerinin önde gelen nedenlerinden biridir ve meme kanseri ölümleri kadınlar arasında yaygındır. Meme kanserinin erken teşhisi, bu ölümleri azaltmanın ana yollarından biri olarak kabul edilir. Uzman sistemler, yapay zekâ (AI) ve makine öğrenmesi (ML) teknikleri, tıp alanındaki doktorlara erken hastalık tespitinde yardımcı olmayı amaçlamaktadır. Bu teknolojilerin temel amaçlarından biri, meme kanseri gibi yaşamı tehdit eden hastalıkları daha erken ve doğru bir şekilde teşhis etmektir. Bu çalışmada, Wisconsin Meme Kanseri Veri Seti (WBCD) analiz edilmiş ve Temel Bileşen Analizi (PCA) ile farklı eksik veri atama yöntemlerinin etkileri değerlendirilmiştir. PCA tabanlı veri indirgeme teknikleri, denetimli ML yöntemlerinde kullanılmaktadır. Bu çalışmada, meme kanseri teşhisi için ML yaklaşımlarının eksik veri yönetimi stratejileriyle birleştirilmesi vurgulanmaktadır. Çalışmanın veri seti 699 veriden oluşmaktadır. Verilerden 16'sı eksik veri olarak tanımlanmıştır. Bu eksik veriler, farklı veri atama yöntemleri kullanılarak işlenmiştir. Medyan tekniğinin en iyi performansı sağladığı görülmüştür. Eksik veriler medyan değer ile doldurulduktan sonra, veri seti üzerinde PCA tabanlı veri indirgeme teknikleri kullanılmıştır. Tümörleri sınıflandırmak için Karar Ağaçları (DT), Doğrusal Regresyon (LR), Lojistik Regresyon (LogR), k En Yakın Komşular (k-NN), Polinom Regresyon (PR), Rastgele Orman (RF) ve Destek Vektör Makineleri (SVM) modellerinin performansları incelenmiştir. Bu tekniklerin etkileri çeşitli PCA bileşen sayılarına sahip ML algoritmaları üzerinde değerlendirilmiştir. En iyi performans SVM ve k-NN algoritmalarında gözlenmiştir. Başarı oranları sırasıyla %97,14 ve %98,57 olarak tespit edilmiştir.

1. INTRODUCTION

Breast cancer is a severe health problem, especially common in women, and its treatability is directly linked to early diagnosis. Early diagnosis can prevent disease progression and increase the chances of successful treatment [1–3].

According to a report published by the World Health Organization (WHO), 20 million new cancer cases occurred worldwide in 2022, resulting in 9.7 million deaths. Breast, colorectal, and lung cancers are among the most common types of cancer globally. It has been reported that 2.3 million breast cancer cases were diagnosed and 685,000 deaths occurred in 2020 [4]. Therefore, it is believed that research and development of early detection methods for breast cancer will play an important role in reducing mortality rates. The rapid development of artificial intelligence (AI) in the last decade shows that significant progress can be made in almost all areas in the future. These developments are in a position to radically change various aspects of human life and our interactions with society.

AI is used for many purposes in almost every field and is divided into different categories. In this context, in the field of medicine especially. AI systems are used effectively and successfully in disease detection and cancer classification. Studies show that ML techniques are essential, especially in medical applications [5,6].

The primary purpose of developing ML and other AI algorithms, as well as other biomedical technologies is to assist doctors in their evaluation processes in healthcare services. All these technologies are used to support doctors' decision-making processes and assist them in issues such as data analysis and image interpretation. The effectiveness and success of ML techniques are significant especially in disease diagnosis and cancer classification. In this context, using AI technologies in medicine provides substantial advantages, such as faster diagnosis of patients and the determination of appropriate treatment methods. In addition, developing and improving these technologies help improve the quality of service in the medical field and the quality of life of patients. Therefore, AI technologies in the medical field are constantly being developed and improved [7,8].

The application of ML techniques to classify of breast cancer has gained increasing attention in various studies. There are different studies on breast cancer diagnosis. A literature review of selected studies is presented for performance comparison and will be discussed here. The WBCD dataset has also been subjected to performance analysis using ML solutions by various researchers. For instance, Hasan et al. [9] used the WBCD and SEER 2017 Breast Cancer Dataset. The model achieved an accuracy rate of 99.1% on the reduced WBCD dataset and 89.3% on the SEER 2017 dataset. In another study, Mushtaq et al. [10] showed the highest accuracy of 99.20% for the Sigmoid-based Naive Bayes method. They also presented that the k Nearest Neighbour (k-NN) method performed the best with PCA-based techniques. The accuracy rate was between 96.4% and 97.8%.

Kong [11] used five different ML models including LR, RF, SVM, k-NN and NB using WBCD. Performance evaluation was performed based on accuracy, precision and recall. The study uses initial tumour data to diagnose breast cancer using ML models. The study presents that the RF model achieved 98.25% prediction accuracy, and the SVM model achieved 100% prediction accuracy. Laghmati et al. [12] discusses the use of PCA and ML algorithms. Particularly the k-NN algorithm, for breast cancer classification and prediction. Sindhuja et al. [13] employed a deep neural network (DNN) incorporating PCA for feature selection to predict breast cancer types and recurrences. So, PCA is used to reduce the size and improve the performance of various ML algorithms. A study showed that integrating PCA with SVM increased the prediction accuracy from 94% to 96% while achieving high precision and recall scores [14]. Another study highlighted the effectiveness of PCA-based Deep Neural Network (PCA-DNN), which achieved an impressive accuracy of 98.83% on the WBCD [15]. Additionally, combining PCA with ensemble methods such as Gradient Boosting further enhances the prediction capabilities, allowing for robust classification of benign and malignant cases [16,17]. However, the need for larger datasets to improve model performance remains a common limitation across studies [16]. A hybrid approach of Convolutional Neural Networks (CNN) and PCA is proposed for early breast cancer diagnosis. This research uses unsupervised PCA for data understanding and supervises CNN for benign/malignant tumour classification from mammography images [18]. PCA is an important tool in optimizing ML models for breast cancer diagnosis [19].

The main objective of this research is to analyse the effects of different missing data imputation methods and PCA data reduction techniques used to improve the performance of ML models for early breast cancer diagnosis. PCA is mainly used as a dimensionality reduction technique that condenses many variables into a more manageable subset while preserving the most relevant information [20]. First, missing data imputation methods are used to appropriately fill in missing or empty values in cases where they exist in the dataset. These methods are essential for maintaining data integrity and ensuring that ML models produce more reliable results. Different missing data imputation methods may employ various strategies, considering the distribution of missing data and the dataset's characteristics. PCA identifies the principal components that emphasize the relationships between variables in the dataset and reduces the dimensionality of the dataset by selecting these components accurately. These results give a more readily understandable dataset. Also, it can improve the performance of ML models. The research detail will directly compare the effects of different missing data imputation methods and PCA-based data reduction techniques on the performance of ML models. This research was conducted on the WBCD [21]. The study's results were analysed using different metrics and

performance criteria to determine which method performed better.

The motivation of this study is to present a different approach to increase the accuracy and reliability of ML models in the diagnosis of breast cancer, which is a major health problem. Despite the advances in diagnostic technologies, the handling of missing data and the high dimensionality of medical datasets may pose varying degrees of challenges. This study emphasizes the importance of appropriate data completion techniques and the PCA method on model performance. It also demonstrates that model performance may be affected if these choices are not considered.

Our main contributions to this study are as follows.

• An analysis of different missing data imputation methods is presented in breast cancer diagnosis

• The effect of PCA on model accuracy was investigated on the new dataset, which was obtained by imputation missing data in the dataset.

• The performance analysis of PCA and missing data completion techniques on the same dataset was evaluated using ML methods.

This paper is organized as follows. Section 2 presents the materials and methods, providing information about the dataset and ML models used in the study. Section 3 presents the experimental results. This section shows the effects of missing data imputation techniques and PCA-based data reduction methods on the model performance. The performance of ML algorithms is compared. The evaluation of experimental results and findings related to other studies are discussed in Section 4. Finally, Section 5 summarizes the study's general conclusions and provides suggestions for future studies.

2. MATERIAL AND METHOD

Machine learning has shown significant advancements in the field of medicine in recent years. These algorithms have developed successful analyses and outcomes for various applications. Early diagnosis, particularly, holds critical importance in the treatment of breast cancer. ML algorithms contribute significantly to health professionals in areas such as early detection of breast cancer, risk analysis, managing treatment processes and decision support.

The analysis of classifiers used in the study was tested with the WBCD dataset. The WBCD dataset consists of 699 samples with 9 features. In the study, out of 9 features, the classification information was used for disease diagnosis, categorized as 4-malignant and 2-benign. Figure 1 shows the distribution of classification information.

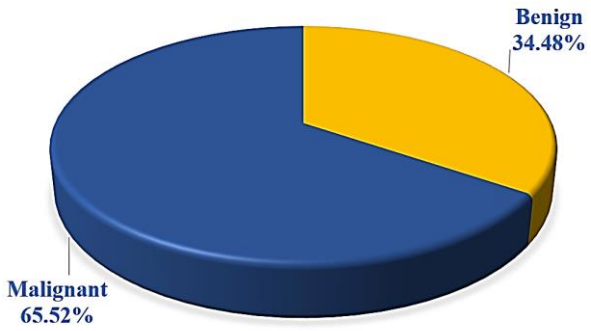

Figure 1. Classification of breast cancer diagnoses

The features used in the classification of tumours are crucial for obtaining the conclusion of whether a tumour is benign or malignant. These features represent various characteristics of the tumour and help determine the characteristics of tumour cells. These features are used to better understand the morphology and behaviour of tumours. ML algorithms can be used to predict whether a tumour is benign or malignant by analysing these features [22]. The dataset is structured as a table containing nine medical parameters and one output class. These parameters encompass various measurements and evaluations obtained during the examination, as illustrated in Table 1.

Table 1. Medical parameters

Specification	Average	Standard Deviation	Minimum Value	Maximum Value
Clump Thickness	4.42	2.82	1.00	10.00
Uniformity of Cell Size	3.13	3.05	1.00	10.00
Uniformity of Cell Shape	3.21	2.97	1.00	10.00
Marginal Adhesion	2.81	2.86	1.00	10.00
Single Epithelial Cell Size	3.22	2.21	1.00	10.00
Bare Nuclei	3.46	3.64	0.00	10.00
Bland Chromatin	3.44	2.44	1.00	10.00
Normal Nucleoli	2.87	3.05	1.00	10.00
Mitoses	1.59	1.72	1.00	10.00
Class	2.69	0.95	2.00	4.00

After the data content was received, an examination was first made to determine whether the data contained discrete and missing values. The analysis revealed that the data did not contain discrete values, but 16 missing data points were observed. Various approaches were explored concerning our studies identified 16 missing data points. The main reason for this decision is that although replacing missing data with a value of 0 has been preferred in previous studies on a similar dataset, the classification of the WBCD dataset has been performed between 1 and 10.

The correlation matrix is a statistical tool that measures the relationship between each pair of variables in a dataset. This relationship indicates how variables change together and how dependent they are on each other. The correlation matrix is presented in matrix form, where each cell contains the correlation coefficient between the corresponding two variables. These coefficients typically range from -1 to 1. As a coefficient approaches 1, the

relationship between variables is stronger and positively oriented; that is, as one variable increases, the other also increases. As it approaches -1, the relationship is negatively oriented; as one variable increases, the other decreases. Coefficients close to 0 indicate little to no relationship between variables, or a very weak relationship. When examining the correlation matrix, attention is paid to high correlation coefficients. High correlation indicates a strong relationship between variables and in such cases, these variables may need to be considered together or one may be preferred over the other in the modelling process. On the other hand, low or near-zero correlation coefficients indicate no or very weak relationship between variables, which helps in identifying unnecessary variables in the modelling process. The correlation matrix for the WBCD dataset is provided in Figure 2.

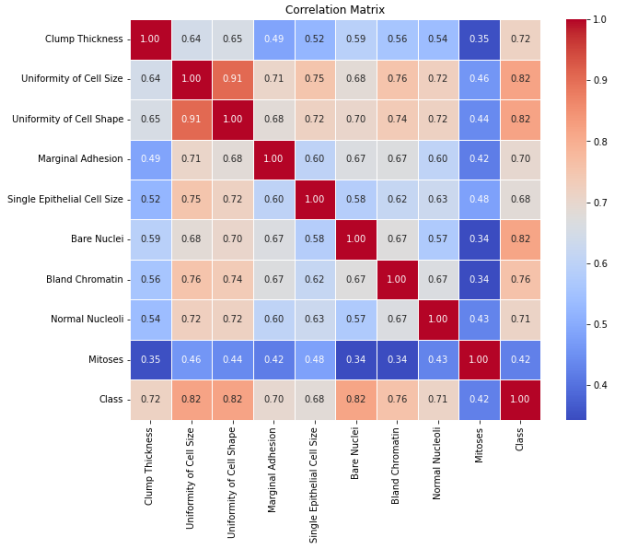

Figure 2. Correlation matrix

Four different methods were used to remove missing data. PCA was used as the dimensionality reduction technique. The obtained results enabled us to compare the effects of different imputation methods and PCA on the model performance. Data processed with PCA after replacing missing values with zero, mean, median values or by deleting missing data were evaluated for model performance. These analyses helped determine which imputation method and PCA component count provided the best performance for the model. In this way, effective handling of missing data and optimization of the model were achieved.

In the study, model performances were measured both across the entire dataset and with lower-dimensional data obtained using the PCA method for feature extraction. During feature extraction, features were transformed into linear combinations of features orthogonal to each other to reduce the dimensionality of the data and improve the accuracy and efficiency of the model. PCA takes the feature set as input and outputs a set of linear combinations of the elements of a subset of the feature set. This two-step approach was implemented as follows: In the first step, the data dimensionality was reduced using a fast and effective unsupervised feature extraction technique like PCA. Subsequently, the obtained lower-dimensional data were used to train the models. Figure 3 depicts the process flow of the proposed method.

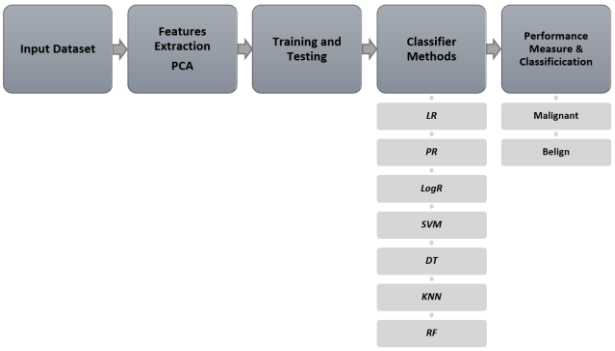

Figure 3. Flowchart of proposed system

ML algorithms exhibit different performances depending on the dataset, the type of problem and the metrics measured. Therefore, it is not always accurate to claim that one algorithm is superior to another. For instance, while one classifier may outperform others in a specific task, a different classifier may yield better results in another task.

In addition, different parameters such as label balance, size, and noise level of the dataset may affect the performance of the algorithms. This study used the Python programming language, Scikit-learn, and TensorFlow libraries within the Anaconda platform using the Spyder environment. The performances of the 16-missing data in the WBDC dataset and the methods of deleting, writing zero, and filling with mean and median values were evaluated. PCA is one of the dimensionality reduction techniques, and the number of components was assessed using LR, PR, LogR, SVM, DT, k-NN and RF algorithms. The obtained results were subjected to a detailed analysis to determine which algorithm performed better and provided the most accurate results for breast cancer diagnosis. It was determined which algorithm provided the best results and most accurate predictions for breast cancer diagnosis with the use of how many components.

3. EXPERIMENTAL RESULTS

In this study, there are a total of 699 data points in the WBCD dataset. with 16 of them containing missing values. Twenty percent of the dataset was set aside to evaluate the model's performance, while the remaining data was used for training the model. Various methods were applied to handle missing data such as zero imputation, mean, median and deletion of missing values. The results of these approaches are presented in Tables 2, 3, 4 and 5 respectively. When considered overall. applying the median imputation method for handling missing data resulted the highest performance among all ML algorithms, as demonstrated in Table 5.

Table 2. Assigning a zero value to replace missing data

Number of Components	LR	PR	LogR	SVM	DT	k-NN	RF
All Components	0.8500	0.8714	0.9643	0.9714	0.9429	0.9786	0.9643
PCA (n=8)	0.8571	0.8857	0.9643	0.9714	0.9357	0.9714	0.9714
PCA (n=7)	0.8500	0.9071	0.9643	0.9714	0.9357	0.9714	0.9714
PCA (n=6)	0.8500	0.9071	0.9571	0.9714	0.9357	0.9643	0.9786
PCA (n=5)	0.8500	0.9000	0.9571	0.9714	0.9429	0.9714	0.9643
PCA (n=4)	0.8571	0.8786	0.9643	0.9714	0.9429	0.9643	0.9714
PCA (n=3)	0.8571	0.8929	0.9714	0.9714	0.9214	0.9714	0.9643
PCA (n=2)	0.8643	0.9143	0.9571	0.9643	0.9429	0.9714	0.9500

Table 3. Assigning mean values instead of missing data

Number of Components	LR	PR	LogR	SVM	DT	k-NN	RF
All Components	0.8500	0.8786	0.9571	0.9643	0.9571	0.9857	0.9643
PCA (n=8)	0.8571	0.8929	0.9643	0.9643	0.9429	0.9714	0.9714
PCA (n=7)	0.8571	0.9000	0.9643	0.9714	0.9214	0.9714	0.9714
PCA (n=6)	0.8500	0.8929	0.9571	0.9714	0.9286	0.9643	0.9786
PCA (n=5)	0.8500	0.9000	0.9571	0.9714	0.9071	0.9714	0.9643
PCA (n=4)	0.8571	0.8786	0.9571	0.9714	0.9286	0.9643	0.9643
PCA (n=3)	0.8571	0.8929	0.9643	0.9714	0.9214	0.9714	0.9714
PCA (n=2)	0.8714	0.9143	0.9643	0.9643	0.9286	0.9714	0.9429

Table 4. Extraction of missing data

Number of Components	LR	PR	LogR	SVM	DT	k-NN	RF
All Components	0.7883	0.8613	0.9562	0.9489	0.9416	0.9489	0.9416
PCA (n=8)	0.7883	0.8613	0.9635	0.9489	0.9781	0.9562	0.9635
PCA (n=7)	0.7883	0.8832	0.9708	0.9562	0.9562	0.9562	0.9854
PCA (n=6)	0.7883	0.8832	0.9489	0.9562	0.9489	0.9562	0.9781
PCA (n=5)	0.7883	0.8686	0.9489	0.9562	0.9343	0.9635	0.9781
PCA (n=4)	0.7883	0.8759	0.9562	0.9562	0.9343	0.9635	0.9708
PCA (n=3)	0.7810	0.8759	0.9562	0.9708	0.9416	0.9781	0.9708
PCA (n=2)	0.7810	0.8978	0.9489	0.9635	0.9489	0.9854	0.9635

Table 5. Filling in missing data with median value

Number of Components	LR	PR	LogR	SVM	DT	k-NN	RF
All Components	0.8571	0.8857	0.9571	0.9714	0.9357	0.9857	0.9643
PCA (n=8)	0.8571	0.8929	0.9643	0.9714	0.9429	0.9714	0.9714
PCA (n=7)	0.8571	0.9071	0.9643	0.9714	0.9429	0.9714	0.9714
PCA (n=6)	0.8500	0.9071	0.9571	0.9714	0.9429	0.9643	0.9643
PCA (n=5)	0.8500	0.9000	0.9571	0.9714	0.9286	0.9714	0.9786
PCA (n=4)	0.8571	0.8786	0.9571	0.9714	0.9429	0.9643	0.9714
PCA (n=3)	0.8571	0.8929	0.9714	0.9714	0.9357	0.9786	0.9714
PCA (n=2)	0.8643	0.9143	0.9571	0.9643	0.9429	0.9714	0.9500

Figure 6 evaluates four methods for handling missing data and seven different ML algorithms. Among these methods, k-NN showed the best performance of 98.57% for mean imputation. In the case of deletion methods, the success rates of the algorithms are generally lower than other methods. The best performance was observed at the median imputation. ML methods using median imputation are 98.57% for k-NN, 97.14% for SVM, and 96.43% for RF. Overall, it was found that deletion or imputation with zeros, with accurate data and PCA applications, led to lower success rates for all ML algorithms. In contrast, imputing missing data with mean or median values improved the overall performance of the algorithms.

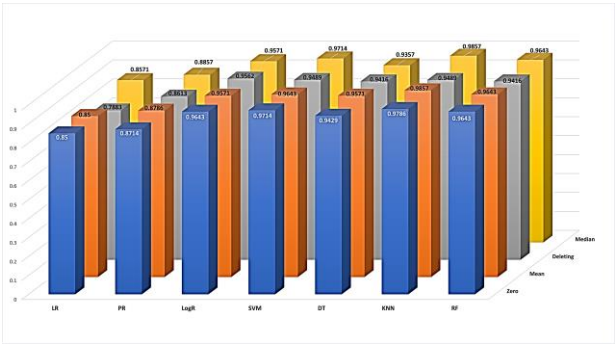

Figure 6. Missing data results for all data sets

4. DISCUSSION

The findings of this study show that missing data management and PCA-based data reduction techniques significantly affect the performance of ML models in breast cancer diagnosis. Our analysis revealed that filling missing data with a median is more effective than other methods in improving model performance. The fact that the median filling method represents the data in a way that is suitable for the distribution in the dataset can be considered one of the main reasons for this effect.

In other studies, the MLTPBC system proposed by Narasimhaiah and Nagaraju [23] uses automatic systems for breast cancer identification using existing methods without PCA. Kong [11] emphasize that PCA achieves 100% in-sample prediction accuracy, especially with the PCA-RF combination, and demonstrates the effectiveness of PCA in classification tasks. Banerjee et al. [24] performed on the same dataset and 16 missing data were deleted. One was trained using Ensemble Learning, and the other without Ensemble Learning. The success rates were given as 95.6% and 82.59%, respectively. While their study was limited to a single method for removing missing data, it was also limited to only two different algorithms in the training part. In their research, Kadhim and Kamil [25] performed it on the same dataset, and 16 missing data were deleted. Eleven different classification models were analysed. The Extreme Random Trees (ERT) model was the most successful, with a success rate of 97.36%. In this study, only one method was used for missing data.

In our study, integrating ML algorithms with PCA with missing data management strategies provided significant improvements in model performance. It has been observed that reducing the data size using PCA leads to significant performance increases, especially in SVM and k-NN algorithms. These results are consistent with the findings obtained in other studies in the literature. It is seen that the combination of median filling and PCA further improves the performance of ML models.

In addition, only ML algorithms were used in [26,27] studies, and breast cancer diagnosis was predicted without using PCA. Unlike these studies, the contribution of the combined use of missing data management and PCA-based data reduction techniques to model performance is seen more clearly in our research. This study emphasizes the importance of using missing data management and PCA-based data reduction techniques to increase the success of ML models in breast cancer diagnosis.

5. CONCLUSION

This study evaluates PCA and ML algorithms with methods for eliminating missing data are used. In the study, the performances of the methods of deleting, writing zero, and filling with mean and median values among the missing data removal methods were evaluated for 16 missing data on the WBDC data set. For each method, the evaluation with different component numbers in the data reduction technique with PCA was tried in 7 different ML algorithms. When the results were analysed, it was observed that the method of deleting missing data generally performed lower. This finding causes low performance in the models because deleting missing data causes data loss. On the other hand, it was observed that the mean imputation method obtained results that were close to those of the median imputation method. Median imputation was effective, especially when the data distribution had a significant asymmetry. However, in some cases, the performance obtained with mean imputation is lower than zero imputation. This variation suggests that the selection of the imputation method may depend on the characteristics of the model and the data distribution. As a result, when the most successful two algorithms among seven were evaluated, the highest performance of the models was observed in filling with median with 97.14% success rates for SVM and 98.57% for k-NN. After filling the missing data with the median filling method, the number of components was evaluated in dimensionality reduction with PCA. The results showed that some models could slightly increase their accuracy rates or maintain their current performance. It was observed that LogR and PR models achieved the highest accuracy rate, especially when the number of components was reduced. However, in SVM and k-NN algorithms, there was no significant change in their success rates even when the number of components was reduced. When working on large data sets, it is thought that reducing the data size will speed up the analysis processes of the data set, which will provide time and cost savings. In future studies, examining the effects of missing data removal methods on large data sets with more missing data will be helpful. In addition to PCA, the effects of dimensionality reduction techniques such as T-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) on model performance can be examined. Evaluating the performance of these methods by applying them to more ML methods and DNN will contribute to obtaining more detailed results.

REFERENCES

[1] Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and preventions of breast cancer. Int J Biol Sci 2017;13:1387–97. https://doi.org/10.7150/ijbs.21635.

[2] Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis,

treatment and follow-up. Annals of Oncology 2019;30:1194–220. https://doi.org/10.1093/annonc/mdz173.
[3] Ginsburg O, Yip CH, Brooks A, Cabanes A, Caleffi M, Yataco JAD, et al. Breast Cancer Early Detection: A Phased Approach to Implementation. Cancer 2020;126:2379–93. https://doi.org/10.1002/cncr.32887.
[4] Global Breast Cancer Initiative Implementation Framework Assessing, strengthening and scaling up services for the early detection and management of breast cancer. Geneva: 2023.
[5] Ting Sim JZ, Fong QW, Huang W, Tan CH. Machine learning in medicine: what clinicians should know. Singapore Med J 2023;64:91–7. https://doi.org/10.11622/smedj.2021054.
[6] Saturi S. Review on Machine Learning Techniques for Medical Data Classification and Disease Diagnosis. Regen Eng Transl Med 2023;9:141–64. https://doi.org/10.1007/s40883-022-00273-y.
[7] Zhang B, Shi H, Wang H. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. J Multidiscip Healthc 2023;16:1779–91. https://doi.org/10.2147/JMDH.S410301.
[8] Savić M, Kurbalija V, Ilić M, Ivanović M, Jakovetić D, Valachis A, et al. The Application of Machine Learning Techniques in Prediction of Quality of Life Features for Cancer Patients. Computer Science and Information Systems 2023;29:381–404. https://doi.org/10.2298/CSIS220227061S.
[9] Hasan MM, Haque MR, Kabir MMJ. Breast Cancer Diagnosis Models Using PCA and Different Neural Network Architectures. 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019, p. 1–4. https://doi.org/10.1109/IC4ME247184.2019.9036627.
[10] Mushtaq Z, Yaqub A, Hassan A, Su SF. Performance Analysis of Supervised Classifiers Using PCA Based Techniques on Breast Cancer. 2019 International Conference on Engineering and Emerging Technologies (ICEET), 2019, p. 1–6. https://doi.org/10.1109/CEET1.2019.8711868.
[11] Kong D. Research on Prediction of Breast Cancer Type using Machine Learning. vol. 2023. 2023.
[12] Laghmati S, Cherradi B, Tmiri A, Daanouni O, Hamida S. Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques. 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), 2020, p. 1–6. https://doi.org/10.1109/CommNet49926.2020.9199633.
[13] Sindhuja M, Poonkuzhali S, Vigneshwaran P. Breast Cancer Classification Model Using Principal Component Analysis and Deep Neural Network. Smart Innovation, Systems and Technologies, vol. 324, Springer Science and Business Media Deutschland GmbH; 2023, p. 137–49. https://doi.org/10.1007/978-981-19-7447-2_13.
[14] Parman NH, Hassan R, Zakaria NH. Breast Cancer Prediction Using Support Vector Machine Ensemble with PCA Feature Selection Method. International Journal of Innovative Computing 2024;14:15–9. https://doi.org/10.11113/ijic.v14n1.461.
[15] Rani P, Kumar R, Jain A, Lamba R, Sachdeva RK, Choudhury T. PCA-DNN: A Novel Deep Neural Network Oriented System for Breast Cancer Classification. EAI Endorsed Trans Pervasive Health Technol 2023;9. https://doi.org/10.4108/eetpht.9.3533.
[16] Lou J. Prediction of breast cancer based on RF, SVM and PCA. J Phys Conf Ser, vol. 2646, Institute of Physics; 2023. https://doi.org/10.1088/1742-6596/2646/1/012036.
[17] Bista C, M A, Slimanzay S, Sheikh MS, Srinivasa Rao P. Breast Cancer Prediction System Utilizing Machine Learning Algorithms, Institute of Electrical and Electronics Engineers (IEEE); 2024, p. 80–4. https://doi.org/10.1109/ieeeconf61558.2024.10585589.
[18] Eladallaoui HE, Elfazziki A, Sadgal M. A CNN and PCA Approach for Earlier Breast Cancer Diagnosis. Proceedings - 17th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2023, Institute of Electrical and Electronics Engineers Inc.; 2023, p. 247–52. https://doi.org/10.1109/SITIS61268.2023.00045.
[19] Yadav RK, Singh P, Kashtriya P. Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. Procedia Comput Sci 2023;218:1434–43. https://doi.org/10.1016/j.procs.2023.01.122.
[20] Bharadiya JP. A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning. Int J Innov Res Sci Eng Technol 2023;8. https://doi.org/10.5281/zenodo.8002436.
[21] Wolberg Wi. Breast Cancer Wisconsin (Original) 1992.
[22] Kaya S, Yağanoğlu M. An Example of Performance Comparison of Supervised Machine Learning Algorithms Before and After PCA and LDA Application: Breast Cancer Detection. 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, p. 1–6. https://doi.org/10.1109/ASYU50717.2020.9259883
[23] Narasimhaiah P, Nagaraju C. Machine Learning Technique for Prediction of Breast Cancer. International Journal on Recent and Innovation Trends in Computing and Communication 2023;11:368–80. https://doi.org/10.17762/ijritcc.v11i7s.7012.
[24] Baneriee C, Paul S, Ghoshal M. A Comparative Study of Different Ensemble Learning Techniques Using Wisconsin Breast Cancer Dataset. 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2017, p. 1–6. https://doi.org/10.1109/ICCECE.2017.8526215.
[25] Kadhim RR, Kamil MY. Comparison of breast cancer classification models on Wisconsin dataset. International Journal of Reconfigurable and Embedded Systems 2022;11:166–74. https://doi.org/10.11591/ijres.v11.i2.pp166-174.

[26] Asharma S, Shinde S, Choudhary A, Raj U, Srivastava P. Machine Learning-Based Breast Cancer Detection. 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2024, Institute of Electrical and Electronics Engineers Inc.; 2024. https://doi.org/10.1109/ICRITO61523.2024.10522209.

[27] Ahmed SS, Srivastava Y, Khan MohdG. Prediction and Diagnosis of Breast Cancer Using Machine Learning Algorithms. Asian Journal of Research in Medical and Pharmaceutical Sciences 2024;13:54–60. https://doi.org/10.9734/ajrimps/2024/v13i3261.