ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 618.19-006:004.942 **doi: 10.32620/reks.2023.4.03**

Sk. Shalauddin KABIR[1], Md. Sabbir AHMMED[2], Md. Moradul SIDDIQUE[1], Romana Rahman EMA[1], Motiur RAHMAN[2], Syed Md. GALIB[1]

[1] *Department of Computer Science and Engineering, Jashore University of Science and Technology, Jashore-7408, Bangladesh*
[2] *Department of Computer Science and Engineering, NUBTK Khulna-9100, Bangladesh*

BREAST TUMOR PREDICTION AND FEATURE IMPORTANCE SCORE FINDING USING MACHINE LEARNING ALGORITHMS

*The **subject matter** of this study is breast tumor prediction and feature importance score finding using machine learning algorithms. The **goal** of this study was to develop an accurate predictive model for identifying breast tumors and determining the importance of various features in the prediction process. The **tasks** undertaken included collecting and preprocessing the Wisconsin Breast Cancer original dataset (WBCD). Dividing the dataset into training and testing sets, training using machine learning algorithms such as Random Forest, Decision Tree (DT), Logistic Regression, Multi-Layer Perceptron, Gradient Boosting Classifier, Gradient Boosting Classifier (GBC), and K-Nearest Neighbors, evaluating the models using performance metrics, and calculating feature importance scores. The **methods** used involve data collection, preprocessing, model training, and evaluation. The **outcomes** showed that the Random Forest model is the most reliable predictor with 98.56 % accuracy. A total of 699 instances were found, and 461 instances were reached using data optimization methods. In addition, we ranked the top features from the dataset by feature importance scores to determine how they affect the classification models. Furthermore, it was subjected to a 10-fold cross-validation process for performance analysis and comparison. The **conclusions** drawn from this study highlight the effectiveness of machine learning algorithms in breast tumor prediction, achieving high accuracy and robust performance metrics. In addition, the analysis of feature importance scores provides valuable insights into the key indicators of breast cancer development. These findings contribute to the field of breast cancer diagnosis and prediction by enhancing early detection and personalized treatment strategies and improving patient outcomes.*

Keywords: *Breast tumor; Benign; Classification model; Machine learning; Tumor; Malignant; Data optimization.*

Introduction

An abnormal mass of tissue is referred to as a tumor. Excessive cell division and growth lead to the formation of tumors. Tumors may be benign or malignant. Benign tumors develop gradually and do not metastasize (spread to other portion of the body), it is not a cancerous tumor. Malignant tumors are abnormal growths of cells that can invade nearby tissues and spread to other parts of the body. Through the lymphatic and blood systems, it can also spread to other bodily areas and is called a neoplasm [1]. Cell division is the process by which human cells develop and reproduce. Cells grow and become old; they die. Again, new cells take their place and continue to work as workers for the human body. However, sometimes their work process breaks down and abnormal cells grow. These cells may turn into tumors, which are lumps of tissue and some are uncontrollable. These uncontrollable cells are called cancerous cells. Cancerous cells are also called malignant tumors [2]. The world is worried about women's breast cancer.

The most common form of cancer in women is breast cancer, which has several molecular characteristics [3]. In 2020, 2.3 million women were affected by breast cancer, with 68,500 deaths. As of the end of 2020, 7.8 million women had been diagnosed with breast cancer in the past 5 years [4]. This makes it the most common cancer on the planet.

Moreover, in developing countries, young women face more problems. To cope with this problem, early detection of tumors is the best way to obtain proper medical treatment. Therefore, we have used modern technology such as machine learning to detect the types of tumors explicitly. Machine learning is a branch of artificial intelligence (AI) that concentrates on using data and algorithms to simulate how people learn, with the aim of progressively increasing accuracy [5 - 7].

We have proposed five machine learning classification algorithms that will help specialists provide proper clinical treatment according to the type of tumor. Here, our focus is to detect breast tumors and breast cancer. To detect it properly, a fine-needle aspiration (FNA) method is used [8]. FNA is a standard method

for testing cancer cells. During FNA, a small amount of breast tissue is taken from a suspicious area using a thin, hollow needle. After that, the tissue is observed under a microscope with 9 quantities very carefully and assigned a number for each quantity between 1 and 10 [9]. They are clump thickness, marginal adhesion, bare nuclei, and uniformity of cell size, bland chromatin, uniformity of cell shape, single epithelial cell size, mitoses, and normal nucleoli. A large number usually indicates a higher chance of cancer. However, a particular measurement cannot determine whether the sample is benign or malignant.

In this study, we used the Wisconsin Breast Cancer original dataset (WBCD) created by Dr. William H. Wolberg at the University of Wisconsin Hospital. We have analyzed and optimized these data according to our model's needs. We used five classification models: Random Forest, Logistic Regression, Multi-Layer Perceptron, K-Nearest Neighbors, and Gradient Boosting Classifier. We used 10-fold cross-validation methods to ensure the model's accuracy performance. We also analyzed training accuracy and testing accuracy to check dataset health for underfit and overfit. We compared all the models' accuracy and selected the best one. The Random Forest model has given us 98.56 % accuracy as well as 10-fold cross-validation scores, which are better than those of the others. The primary aims of our study are as follows:

– to the proper use of machine learning algorithms for breast cancer early detection;

– the cost of time for the test will hopefully be reduced;

– to obtain the highest accuracy, we evaluated and optimized the WBCD dataset;

– we have shown different machine learning algorithms and compared them. In addition, analytical data visualization is another purpose.

1. Literature Review

S. Ara et al. [10] proposed a machine learning-based model for predicting breast cancer using the WBCD diagnostic dataset. The dataset was obtained from the UCI machine learning repository. There were 569 incidents, 357 of which were benign and 212 were malignant. They proposed several machine learning models: Support Vector Machine, Logistic Regression, K- Nearest Neighbor, Decision Tree, Naive Bays, and Random Forest classifiers. From these models, Random Forest and Support Vector Machine gave outstanding results with 96.5 % accuracy. V. Chaurasia et al. [11] used data mining techniques to predict benign and malignant tumors. They used data from the UCI repository, which had 699 instances, 2 classes (malignant and benign), and 9 integer-valued features. They analyzed the data using the Waikato Environment for Knowledge Analysis (WEKA) tools version 3.6.9. They applied 10-fold cross- validation methods to measure the unbiased estimates of the three popular data mining algorithms: Naive Bayes, RBF Network, and J48 Decision Tree. According to the results, Naive Bayes 97.36 %, RBF Network 96.77 %, and J48 came out with 93.41 %. Y. Li et al. [12] evaluated the performance of machine learning methods for breast cancer. They used two datasets: the Breast Cancer Coimbra Dataset (BCCD) and WBCD. The BCCD contains 116 instances with 10 attributes for each case that was created by M. Patricio et al. [13] at the Faculty of Medicine of the University of Coimbra, and the WBCD involves 699 instances with 10 attributes. They applied five different classification models: Decision Tree, Random Forest, Support Vector Machine, Neural Network, and Logistics Regression. From those models, the Random Forest gave the best results for BCCD with 74.3 % accuracy, 78 % F-measure metric, 78.5 % AUC, and for WBCD with 96.1 % accuracy, 95.5 % F-measure metric, and the AUC score of 98.9 %. H. Asri et al. [14] used machine learning algorithms to predict breast cancer risk and diagnose it. They proposed different algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes, and K-nearest Neighbors on the Wisconsin Breast Cancer (original) dataset. The dataset contains 65.5 % malignant and 34.5 % benign with 11 integer-valued attributes. The SVM gave them the highest accuracy (97.13 %) with the lowest error rate. All experiments were carried out using the WEKA data mining tool within a simulation environment. As technology continues to advance, machine learning will likely play an increasingly important role in improving healthcare outcomes and reducing the burden on healthcare systems.

In conclusion, the integration of machine learning into breast cancer detection has the potential to revolutionize the field by enhancing accuracy, reducing false positives, and enabling more personalized risk assessments. This literature review highlights the growing relevance of machine learning in breast cancer detection and sets the stage for the subsequent sections of this research paper, which delve into the methodology, findings, and implications of the study.

2. Methodology

The proposed Brest Tumor classification methodology is shown in Fig. 1. We divide this procedure into several subsections for data collection, data cleaning, data preprocessing, data analysis, data splitting into training and testing, and evaluation results using various machine learning models.

– Dataset Description;

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

- Data Cleaning and Preprocessing;
- Dataset Analysis;
- Machine Learning for Classification;
- Model Evaluation.

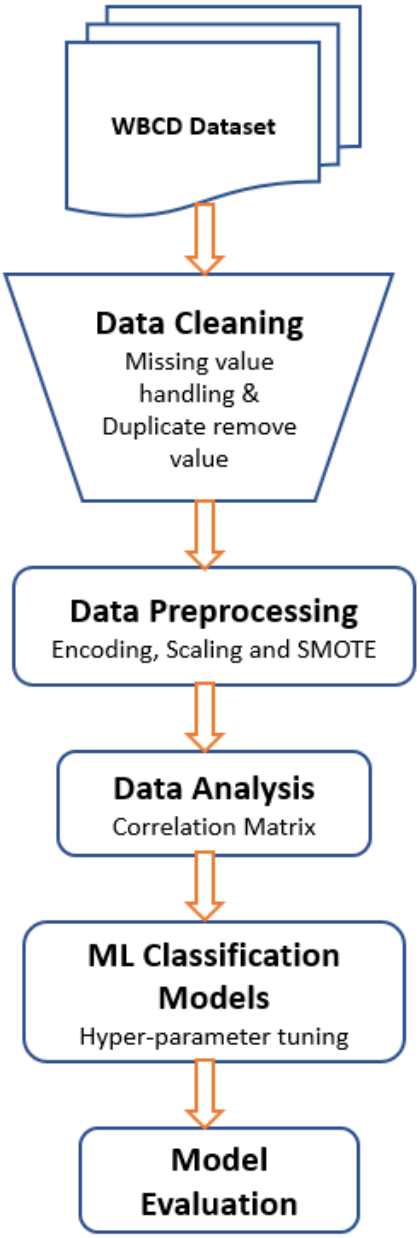

Fig. 1. Flowchart of our proposed Brest Tumor classification methodology

2.1. Dataset Description

In this study, we used the" Breast Cancer Wisconsin (original) Dataset" from the UCI machine learning repository [15], which is publicly accessible. This dataset contains 699 instances and 10 attributes, of which 458 are benign and 241 are malignant, as shown in Fig. 2. Attributes: Clump Thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epi Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class. Here, except for the Class column, all are features. The Class attribute contains binary nominal values (benign and malignant), and each feature contains integer values between 1 and 10. Usually, a large integer number indicates a high chance of malignancy.

Fig. 2. Distribution of Benign and Malignant in the dataset

During the fine-needle aspiration (FNA) method, a small amount of breast tissue is taken from a suspicious area using a thin, hollow needle. After that, the tissue is observed under a microscope with 9 quantities very carefully and assigned a number for each quantity between 1 and 10. The 9 real-value features are described in Table 1.

2.2. Data Cleaning and Preprocessing

In this dataset, we found some missing values defined by '?'. These NaN (Not a Number) values are handled by the Backward Filling technique (BFill). We have tried some other methods too, for example: mean, mode, median, min, max, FFill (Forward Filling) and by the interpolation value. However, by applying the BFill method, we have obtained outstanding results compared to others. BFill() is used to backfill the dataset's missing values. NaN values in the pandas dataframe will be retroactively filled in [17, 23].

This dataset also contains some duplicate values, which we managed by removing the duplicate values from our dataframe. Here, we used the pandas library to handle missing values and remove duplicate values. After removing some data, we obtained some changes in our dataset. Before removing duplicate values, we identified 458 benign and 241 malignant instances. However, after removing duplicate values, we achieved 223 benign and 238 malignant instances, as shown in Figure 4. We are noticing that there is a huge change in benign, as shown in Fig. 3. Here, 0 means benign and 1 means malignant.

Table 1

Information on the features of our dataset

Feature Name	Description
Bare_Nuclei	It refers to nuclei that are not encircled by the cytoplasm (the cell's interior). In benign tumors, they are consistently observed.
Clump_Thickness	In terms of clump thickness, cancerous cells tend to form multilayer clumps, whereas benign cells typically form monolayer clumps.
Cell_Shape/ Size_Uniformity	Cancer cells typically differ from normal cells in size and shape consistency.
Normal_Nuclcoli	The nucleus contains tiny structures known as normal nucleoli. In normal cells, the nucleolus is usually quite tiny, if detectable.
Bland_Chromatin	Bold chromatin refers to the constant 'texture' of the nucleus in healthy cells. Chromatin is typically more agglomerated in cancer cells.
Marginal_Adhesion	Normal cells often adhere to one another in the event of marginal adhesion; however, malignant cells typically lose this ability. Therefore, a lack of adhesion is an indication of cancer.
Mitosis	This is the process through which cells divide and multiply. By counting the mitoses, pathologists can assess the cancer grade.
Single_Epi_Cell_Size	Epithelial cells are determined by both in terms of both shape and layer number. Epithelial cells that are simple have only one layer. Therefore, we can make a decision after observing the size of epithelial cells [16].

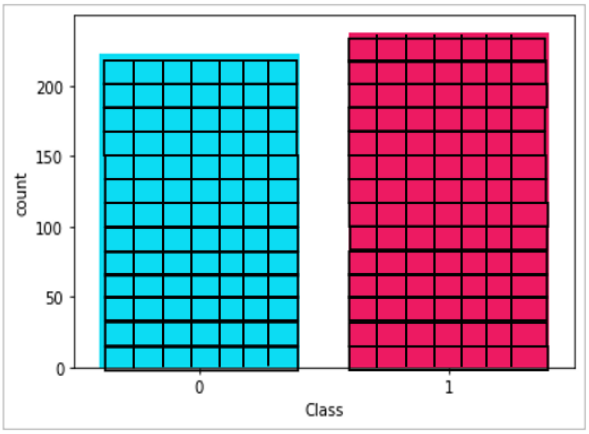

Fig. 3. Distribution of the dataset after removing duplicates

Thus, at this stage, our total instances are now 461. However, this data frame is still unbalanced. To solve this problem, we applied the Synthetic Minority Oversampling Technique (SMOTE). Therefore, for this, the instances of benign and malignant are now equal and both are 238 separately. We also noticed that our target column's (Class) values are string or binary nominal values. However, we need numeric values instead of strings because our machine learning models expect numeric values for better performance. This is why we need to encode the Class attribute. For this, the LabelEncoder algorithm has been used. The LabelEncoder class comes from the preprocessing of the sklearn package. After encoding, we obtained numeric values, where 0 refers to benign and 1 refers to malignant.

Some machine learning algorithms are performing well after feature scaling. Mainly, Neural Network, K-Nearest Neighbors, and Gradient Boosting Classifier are giving us better results after scaling. Therefore, we have applied the feature normalization method to our dataset, except for the Class attribute. We used MinMaxScaler from the preprocessing of the sklearn package. The MinMaxScaler is a scaling technique in which values are shifted and reshaped into a particular range. By default, a new reshaped value takes a number within the 0 to 1 range, and we have also used this range in this study. The formula of MinMaxScaler is shown in Equation (1)

$$X(new) = \frac{X(i) - X(min)}{X(max) - X(min)}, \quad (1)$$

here, X(new) is a new reshaped value. On the other hand, X(i) is a value that we want to reshape. X(min) is a small value of that attribute or column and X(max) is a large value of that column. At this stage, we have performed the encoding and scaling procedure. Now we have applied the SMOTE technique that we have already mentioned. SMOTE is a procedure that can help us balance data for an unbalanced dataset. For this, we have used SMOTE from the oversampling of the imblearn package.

2.3. Dataset Analysis

Data analysis is a vital part before model fitting, whereas different types of decisions are made by applying statistical analysis. In terms of Correlation analysis,

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

the input variables can help us to find different relationships between two features. This connection assists us in determining which input variables are more crucial for the dependent variables. We can accurately predict the outcome of a dependent variable. The correlation coefficient is calculated in the range between -1 and +1. Whereas, a score close to +1, however, indicates a significant positive correlation. On the other hand, a score close to -1 indicates a significant negative correlation [18]. From Fig. 4, it can be seen that the correlation coefficient has the highest value of 0.88, which means it has no strong correlation with others. Although this value is considerable, we did not eliminate any features from our dataset.

We also analyzed the importance of the features, as shown in Fig. 5. We see that Bare Nuclei are the most important feature for our dataset. Additionally, uniform cell size and shape have an impact on this dataset, and their correlation coefficient is 0.88. Therefore, for this aspect of importance, we did not eliminate any features from our dataset.

2.4. Machine Learning for Classification

Before applying machine learning models, the dataset needs to be initially divided into testing and training sets. For this reason, the dataset was split into 30 % of the data for testing and 70 % of the data for training.

A subfield of artificial intelligence and computer science called "machine learning" employs algorithms that are intended to learn from previous learning that can be used to predict the future. It is an automatic procedure. The machine learning model learns in almost the same way that humans learn and gradually improves its accuracy. We used the following algorithms for this work: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Gaussian Naïve Bayes (GaussianNB), Multi-Layer Perceptron (MLP), Gradient Boosting Classifier (GBC).

To execute machine learning models, we used the Python programming language, and sklearn (scikit-learn) is a package that provides us with various machine learning algorithms. After model fitting, we evaluated every model and used 10-fold cross-validation for performance analysis and comparison. After analyzing the results, we applied the hyperparameter tuning method (RandomizedSearchCV or GridSearchCV) as needed.

3. Results and Analysis

In this section, we assess the algorithm's effectiveness on the dataset after implementing machine learning models. Additionally, we measured the implemented system's performance based on accuracy, precision, recall, f1-score, and AUC score from the ROC curve.

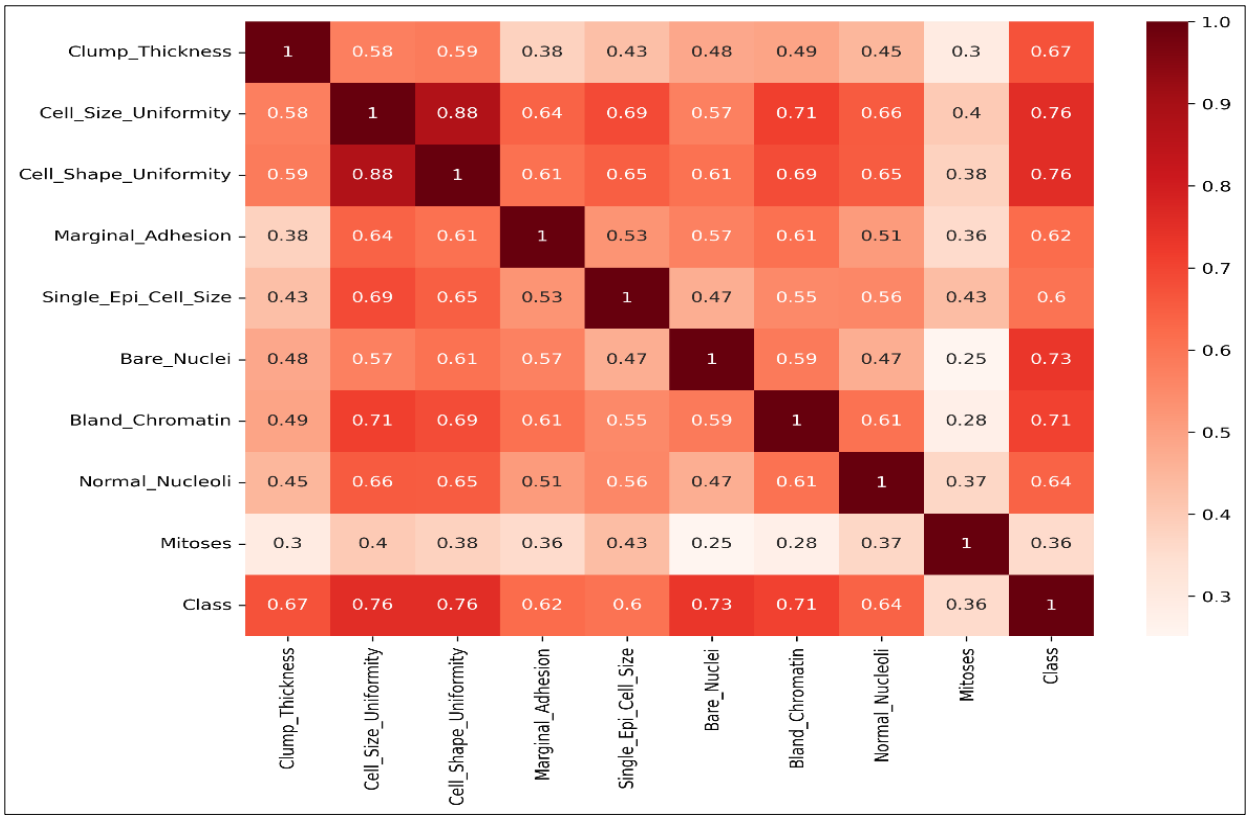

Fig. 4. Correlation among the input variables in the dataset

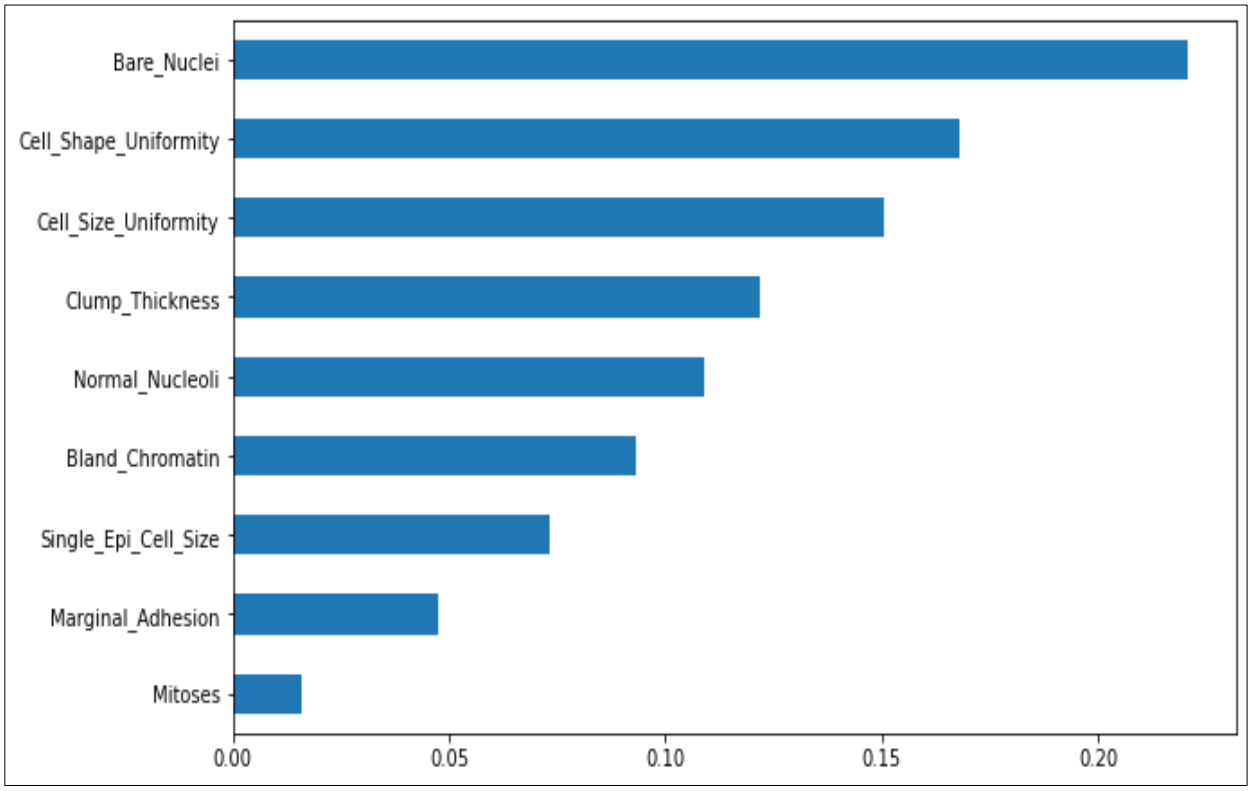

Fig. 5. Feature Importance score of the dataset

According to our findings in Fig. 6, the model accomplished a classification accuracy of approximately 98.56 % with a Precision score 98.57 %, Recall score 98.57 % and F1 score 98.57 %. The Decision Tree (DT) model accomplished a classification accuracy of approximately 89.93 % with a Precision score 95.16 %, Recall score 84.29 % and F1 score 89.39 %. The Logistic Regression (LR) model accomplished a classification accuracy of approximately 97.12 % with a Precision score 98.53 %, Recall score 95.71 % and F1 score 97.10 %. The Multi-Layer Perceptron (MLP) model accomplished a classification accuracy of approximately 97.84 % with a Precision score 98.55 %, Recall score 97.14 % and F1 score 97.84 %. The Gaussian Naïve Bayes (GaussianNB) model accomplished a classification accuracy of approximately 95.68 % with a Precision score 95.71 %, Recall score 95.71 % and F1 score 95.71 %.

The Gradient Boosting Classifier (GBC) model accomplished a classification accuracy of approximately 98.56 % with a Precision score 98.57 %, Recall score 98.57 % and F1 score 98.57 % which are same result of Random Forest Model. Finally, the K-Nearest Neighbors (KNN) model accomplished a classification accuracy of approximately 97.84 % with a Precision score 98.55 %, Recall score 97.14 % and F1 score 97.84 %.

Also, Fig. 6 presents the classification reports of the models as well as the confusion matrix (A), the classification reports (B) of the models, and the ROC curve (C) of the models, where the AUC score from the ROC curve is 99 %, 90 %, 100 %, 99 %, 98 %, 99 % and 98 % for RF, DT, LR, MLP, GaussianNB, GBC and KNN respectively.

4. Discussions

The findings can be compared from Table 2 when the execution of all seven ML techniques was used to identify breast cancer. We achieved the same results for RF and GBC for accuracy, precision, recall, f1-score, and AUC, which are respectively 98.56 %, 98.57 %, 98.57 %, 98.57 %, and 99 %. On the other hand, for Multi-Layer Perceptron and K-Nearest Neighbors, both have gained 97.84 % accuracy, and their AUC results are 99 % and 98 %, respectively. The Logistic Regression gives us 97.12 % accuracy and a 100 % AUC result.

After analysis of Table 2, the Random Forest and the Gradient Boosting Classifier have performed outstandingly. After analysis of 10-fold cross-validation, we noticed that the Random Forest model performs better than the Gradient Boosting Classifier as shown Fig. 7. Therefore, for this as the best model, we have selected Random Forest. Not only that, but we are also trying to determine the dataset's health status [19, 20]. Is it an under-fit or an over-fit? as shown in Fig. 8. For this case, we used the Random Forest classification and analyzed the training and testing results. We have seen that the maximum distance between training and

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

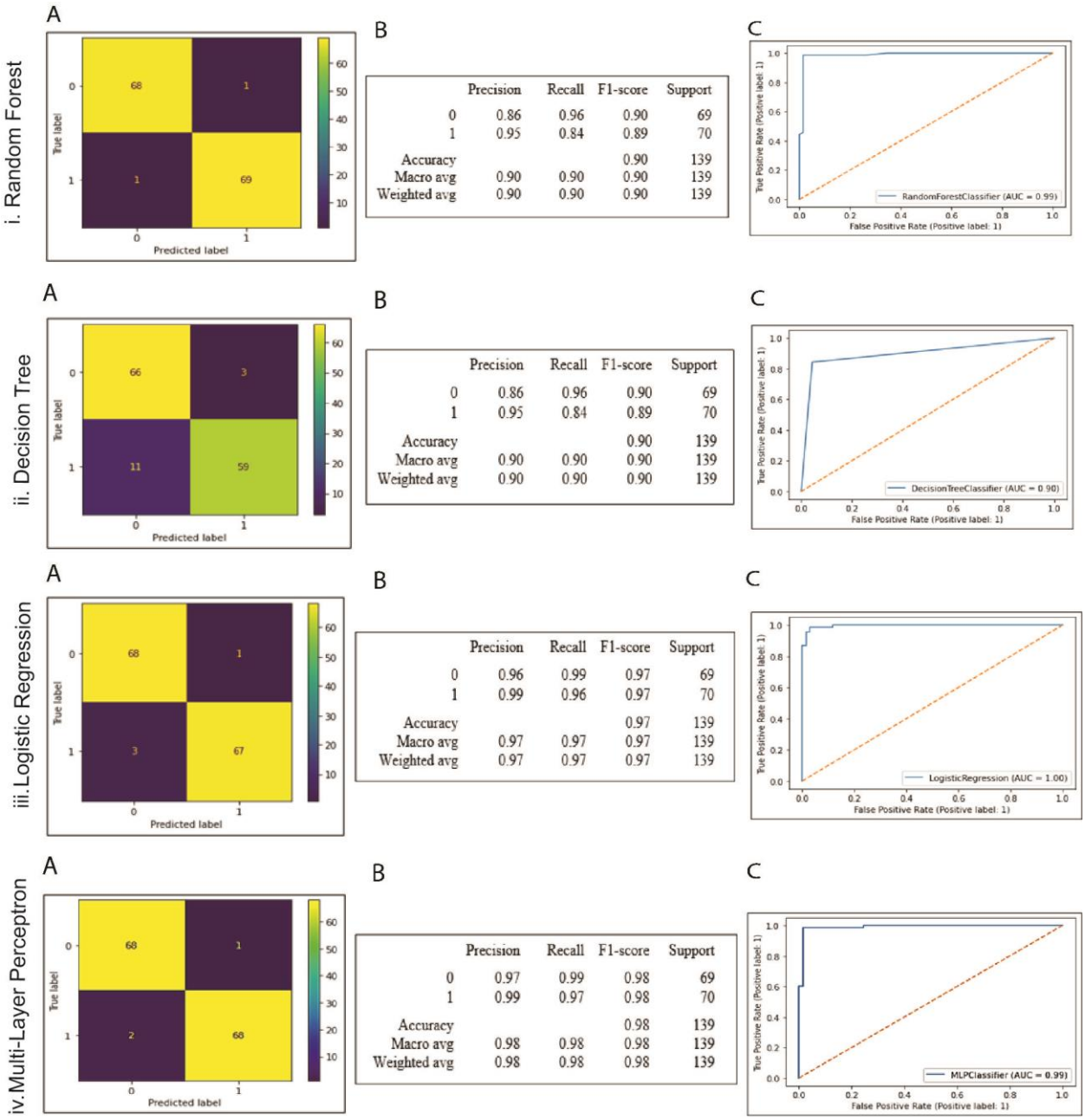

Fig. 6. Confusion Matrix (A), classification re-ports (B) and ROC curve with AUC (C) of the i. Random Forest, ii. Decision Tree, iii. Logistic Regression (LR), iv. Gaussian Naïve Bayes

testing ac-curacy is 6 % or something like that, and the mini-mum distance we found is 1.4 % or something like that. Therefore, we can say that our dataset is good-fitting. Both the training and testing steps have pro-gressed well. There are different studies in the literature due to the emergence of Breast Cancer disease. Here, we compared our proposed model with other approach-es, and the results are given in Table 3.

Conclusions

The main goal of the study is to detect breast tu-mors early so that doctors can easily prescribe and cure them. In this case, we achieved a successful outcome by using several machine learning techniques. We also evaluated these algorithms and applied 10 cross-validations and optimization techniques. As a result, applying Random Forest and Gradient Boosting classi-fication, achieved outstanding outcomes.

By analyzing some facts, we have accepted the Random Forest model as the final model selection. We achieved 98.56 % accuracy and 99 % AUC score from the ROC curve. We also checked the dataset health by observing the training and testing accuracy results. We did not observe over- or under-fitting behavior.

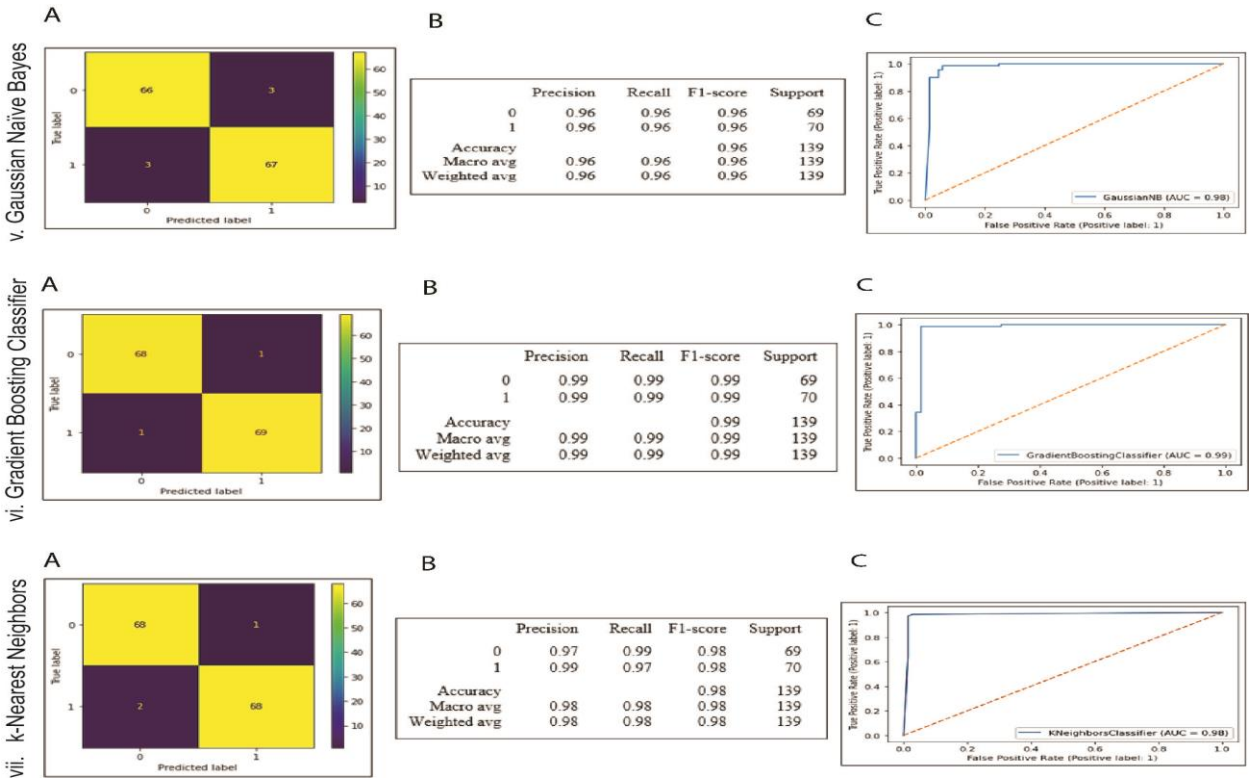

Fig. 6. Confusion Matrix (A), classification re-ports (B) and ROC curve with AUC (C)
of the v. Multi-Layer Perceptron, vi. Gradient Boosting Classifier, vii. K-Nearest Neighbors model

Table 2

Comparing the Results of ML Models

Algorithm	Accuracy, %	Precision, %	Recall, %	F1 Score, %	AUC, %
Random Forest (RF)	98.56	98.57	98.57	98.57	99
Decision Tree (DT)	89.93	95.16	84.29	89.39	90
Logistic Regression (LR)	97.12	98.53	95.71	97.10	100
Multi-Layer Perceptron (MLP)	97.84	98.55	97.14	97.84	99
Gaussian Naïve Bayes (GaussianNB)	95.68	95.71	95.71	95.71	98
Gradient Boosting Classifier (GBC)	98.56	98.57	98.57	98.57	99
K-Nearest Neighbors (KNN)	97.84	98.55	97.14	97.84	98

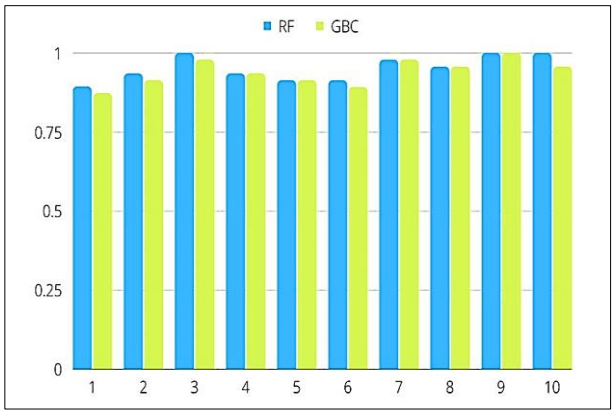

Fig. 7. 10-fold cross-validation results
of RF and GD classifier

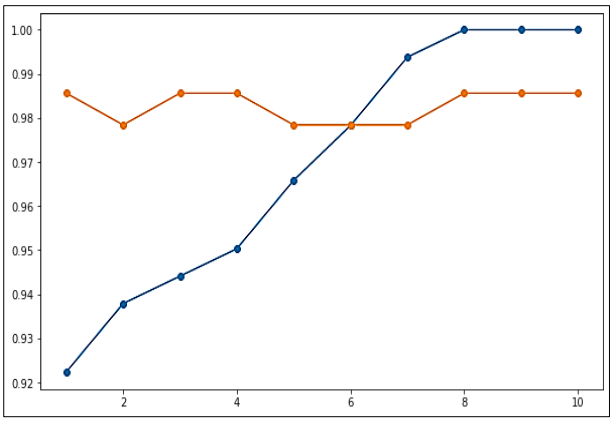

Fig. 8. Dataset Health Status Using RF Classifier

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Table 3

Comparison and Benchmark Studies of the Proposed Method

References	Number of Samples	Model	Accuracy, %	Precision, %	Recall, %	F1-Score, %
Y. Li, *et al.* [12]	683(benign=458, malignant=241)	RF	96.1			95.5
V. Chaurasia, *et al.* [11]	683(benign=458, malignant=241)	Naïve Bayes	97.36		97.4	
S. Ara, *et al.* [10]	569(benign=357, malignant=212)	RF, SVM	96.5			
H. Asri, *et al.* [14]	699(benign=458, malignant=241)	SVM	97.13	98	96	95
V. Chaurasia, *et al.* [11]	683(benign=458, malignant=241)	SMO	96.2	94.6	94.6	
Proposed Method	**461(benign=223, malignant=238)**	**RF**	**98.56**	**98.57**	**98.57**	**98.57**

Our dataset was treated as a good fit dataset. We want to again mention that we have worked on two types of tumors, where 0 means benign and 1 means malignant tumors.

Although our current models have shown improved results, there are still limitations that need to be addressed. One way to enhance the model's performance is by increasing the amount of data and incorporating more features. In this study, we focused on 10 features that could be expanded upon in future investigations. It is advisable to gather local data to supplement the training and testing datasets, as this can provide valuable insights and improve the model's applicability to specific contexts. Another potential avenue for boosting model performance is to leverage deep learning techniques, specifically tumor image segmentation. This involves the use of advanced algorithms, such as convolutional neural networks (CNNs) [21, 22], to identify and isolate tumor regions within medical images. By exploring this approach, we have the opportunity to enhance the accuracy and precision of tumor detection and characterization, which can greatly impact cancer diagnosis, treatment planning, and patient monitoring. However, implementing deep learning techniques and tumor image segmentation requires substantial computational resources, specialized expertise, and a large annotated dataset. Therefore, careful consideration should be given to these factors before embarking on such research endeavors.

In the near future, integrating early detection methods with treatment planning to streamline the transition from diagnosis to personalized treatment strategies will be applied. In addition, advanced imaging technologies, such as improved mammography, are needed to enhance early detection and reduce false positives and false negatives. Moreover, we can employ advanced technology such as neural networks [24].

Author contributions: coding, conceptualization, comparison and original drafting of the manuscript were performed – **Sk. Shalauddin Kabir;** data curation, Methodology, original drafting and review of the manuscript were performed – **Md. Sabbir Ahmmed;** methodology, validation, Investigation and data curation of the manuscript were performed – **Md. Moradul Siddique;** guidance, review, analysis, comparison and drafting of the manuscript were also performed – **Romana Rahman Ema;** comparative analysis, and drafting of the manuscript were performed – **Motiur Rahman;** validation, investigation, review and supervision of the manuscript were performed – **Syed Md. Galib**.

All the authors have read and agreed to the published version of this manuscript.

References

1. *Definition of tumor, NCI Dictionary of Cancer Terms*. Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor (accessed: Feb. 23, 2023).

2. *What Is Cancer?* Available at: https://www.cancer.gov/about-cancer/understanding/what-is-cancer/ (accessed October 11, 2021).

3. Testa, U., Castelli, G., & Pelosi, E. Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments. *Medical Sciences*, 2020, vol. 8, no. 1, article no. 18. DOI: 10.3390/medsci8010018.

4. *Breast Cancer Facts and Statistics.* Available at: https://www.breastcancer.org/facts-statistics (Accessed on Jan. 19, 2023).

5. Gayathri, B. M., Sumathi, C. P., & Santhanam, T. Breast cancer diagnosis using machine learning algorithms – a survey. *International Journal of Distributed and Parallel Systems (IJDPS)*, 2013, vol. 4, iss. 3, pp. 105-112. DOI: 10.5121/ijdps.2013.4309.

6. Nemade, V., Pathak, S., & Dubey, A. K. A systematic literature review of breast cancer diagnosis using machine intelligence techniques. *Archives of Computational Methods in Engineering*, 2022, vol. 29, no. 6, pp. 4401-4430. DOI: 10.1007/s11831-022-09738-3.

7. Elsadig, M. A., Altigani, A., & Elshoush, H. T. Breast cancer detection using machine learning approaches: a comparative study. *International Journal of Electrical & Computer Engineering*, 2023, vol. 13, no. 1, pp. 736-745. DOI: 10.11591/ijece.v13i1.pp736-745.

8. Mangasarian, O. L., & Wolberg, W. H. *Cancer diagnosis via linear programming.* University of Wisconsin-Madison. Computer Sciences Department, 1990. 5 p. Available at: http://digital.library.wisc. edu/1793/59346. (Accessed on Dec. 23, 2022).

9. Lee, H., Yoon, T. J., Figueiredo, J. L., Swirski, F. K., & Weissleder, R. Rapid detection and profiling of cancer cells in fine-needle aspirates. *Proceedings of the National Academy of Sciences*, 2009, vol. 106, no. 30, pp. 12459-12464. DOI: 10.1073/pnas.0902365106.

10. Ara, S., Das, A., & Dey, A. Malignant and benign breast cancer classification using machine learning algorithms. In *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2022, pp. 97-101. DOI: 10.1109/ICAI52203.2021.9445249.

11. Chaurasia, V., Pal, S., & Tiwari, B. B. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, 2018, vol. 12, no. 2, pp. 119-126. DOI: 10.1177/1748301818756225.

12. Li, Y., & Chen, Z. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math*, 2018, vol. 7, no. 4, pp. 212-216. DOI: 10.11648/j.acm.20180704.15.

13. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, 2018, vol. 18, no. 1, article no. 29, pp. 1-8. DOI: 10.1186/s12885-017-3877-1.

14. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 2016, vol. 83, pp. 1064-1069. DOI: 10.1016/j.procs.2016.04.224.

15. Wolberg, W. Breast Cancer Wisconsin (Original). Dataset. *UCI Machine Learning Repository*, 1992. DOI: 10.24432/C5HP4Z.

16. Kurn, H., & Daly, D. T. *Histology, epithelial cell, StatPearls - NCBI BookShelf.* Available at: https://www.ncbi.nlm.nih.gov/books/NBK559063/ (Accessed on Feb. 17, 2023).

17. *What is ffill and bfill in pandas?* Available at: https://www.projectpro.io/recipes/what-is-ffill-and-bfill-pandas (Accessed on Dec. 23, 2022).

18. Kumar, S., & Chong, I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *International journal of environmental research and public health*, 2018, vol. 15, no. 12, article no. 2907. DOI: 10.3390/ijerph15122907.

19. Pothuganti, S. Review on over-fitting and under-fitting problems in Machine Learning and solutions. *Int. J. Adv. Res. Electr. Electron. Instrumentation Eng*, 2018 vol. 7, no. 9, pp. 3692-3695. Available at: http://www.ijareeie.com/upload/2018/september/11A_PS_NC.PDF. (Accessed on Feb. 17, 2023). DOI: 10.15662/IJAREEIE.2018.0709015.

20. Montesinos López, O. A., Montesinos López, A., & Crossa, J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. *In Multivariate statistical machine learning methods for genomic prediction*, 2022, pp. 109-139. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-89010-0_4.

21. Martyniuk, T., Krukivskyi, B., Kupershtein, L., & Lukichov, V. Neural Network model of heteroassociative memory for the classification task. *Radioelectronic and Computer Systems*, 2022, vol. 2, pp. 108-117. DOI: 10.32620/reks.2022.2.09.

22. Krivtsov, S., Meniailov, I., Bazilevych, K., & Chumachenko, D. Predictive model of COVID-19 epidemic process based on neural network. *Radioelectronic and Computer Systems*, 2022, vol. 4, pp. 7-18. DOI: 10.32620/reks.2022.4.01.

23. Tarle, B., & Akkalaksmi, M., Improving classification performance of neuro fuzzy classifier by imputing missing data. *International Journal of Computing*, 2019, vol. 18, iss. 4, pp. 495-501. DOI: 10.47839/ijc.18.4.1619.

24. Striuk, O., & Kondratenko, Yu. Generative adversarial neural networks and deep learning: successful cases and advanced approaches. *International Journal of Computing*, 2021, vol. 20, iss. 3, pp. 339-349. DOI: 10.47839/ijc.20.3.2278.

Received 08.06.2022, Accepted 20.11.2023

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

ПРОГНОЗУВАННЯ ПУХЛИНИ МОЛОЧНОЇ ЗАЛІЗИ І ВИЗНАЧЕННЯ ОЦІНКИ ВАЖЛИВОСТІ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Ск. Шалауддін Кабір, Мд. Саббір Ахммед, Мд. Морадул Сиддік, Романа Рахман Ема, Мотіур Рахман, Саїд Мд. Галіб

Предметом цієї статті є прогнозування пухлин молочної залози та визначення оцінки важливості ознак за допомогою алгоритмів машинного навчання. Метою цього дослідження є розробка точної прогностичної моделі для виявлення пухлин молочної залози та визначення важливості різних ознак у процесі прогнозування. Виконувані завдання включають збір і попередню обробку вихідного набору даних про рак молочної залози Вісконсіна (WBCD). Розподіл набору даних на набори для навчання та тестування, навчання з використанням алгоритмів машинного навчання, таких як випадковий ліс, дерево рішень (DT), логістична регресія, багатошаровий перцептрон, класифікатор посилення градієнта, класифікатор підвищення градієнта (GBC) і K-найближчі сусіди, оцінка моделі з використанням показників ефективності та обчислення балів важливості функцій. Використовувані методи включають збір даних, попередню обробку, навчання моделі та оцінку. Результати показали, що модель Random Forest є найнадійнішим прогностичним фактором із точністю 98,56%. Спочатку ми знайшли 699 екземплярів. Після використання методів оптимізації даних ми досягли 461 екземпляра. Крім того, ми ранжували найкращі функції з набору даних за балами важливості ознак, щоб побачити, як вони впливають на моделі класифікації. Ми використали методи перехресної перевірки (10-кратної) кожної моделі для аналізу продуктивності та порівняння. Висновки, зроблені в результаті цього дослідження, підкреслюють ефективність алгоритмів машинного навчання в прогнозуванні пухлин молочної залози, досягаючи високої точності та надійних показників ефективності. Крім того, аналіз показників важливості ознак дає цінну інформацію про ключові показники розвитку раку молочної залози. Ці висновки сприяють діагностиці та прогнозуванню раку молочної залози, покращують раннє виявлення та персоналізовані стратегії лікування, а також покращують результати пацієнтів.

Ключові слова: пухлина молочної залози; доброякісний; модель класифікації; машинне навчання; пухлина; злоякісний; оптимізація набору даних.

Ск. Шалауддін Кабір – викл. каф. інформатики та інженерії, Науково-технічний університет Джашор, Бангладеш.

Мд. Саббір Ахммед – студент бакалавра факультету інформатики та інженерії, Північний університет бізнесу та технологій Кхулна, Бангладеш.

Мд. Морадул Сиддік – магістрант каф. комп'ютерних наук та інженерії, Джашорський університет науки та технологій, Бангладеш.

Романа Рахман Ема – доц. каф. комп'ютерних наук та інженерії, Джашорський науково-технічний університет, Бангладеш.

Мотіур Рахман – студент бакалавра факультету інформатики та інженерії, Північний університет бізнесу та технологій Кхулна, Бангладеш.

Саїд Мд. Галіб – проф. каф. комп'ютерних наук та інженерії, Джашорський університет науки і технологій, Бангладеш.

Sk. Shalauddin Kabir – Lecturer at Computer Science and Engineering Department, Jashore University of Science and Technology, Bangladesh,
e-mail: sks.kabir@just.edu.bd, ORCID: 0000-0002-0031-8807.

Md. Sabbir Ahmmed – Bachelor's Student at Computer Science and Engineering Department, Northern University of Business and Technology Khulna, Bangladesh,
e-mail: sabbir.cse@yahoo.com, ORCID: 0009-0001-3048-3440.

Md. Moradul Siddique – Masters Student at Computer Science and Engineering Department, Jashore University of Science and Technology, Bangladesh,
e-mail: moradul@just.edu.bd, ORCID: 0000-0003-3264-5383.

Romana Rahman Ema – Assistant Professor at Computer Science and Engineering Department, Jashore University of Science and Technology, Bangladesh,
e-mail: rr.ema@just.edu.bd, ORCID: 0000-0002-2384-9539.

Motiur Rahman – Bachelor's Student at Computer Science and Engineering Department, Northern University of Business and Technology Khulna, Bangladesh,
e-mail: motiurr503@gmail.com, ORCID: 0009-0007-5345-9818.

Syed Md. Galib – Professor at Computer Science and Engineering Department, Jashore University of Science and Technology, Bangladesh,
e-mail: galib.cse@just.edu.bd, ORCID: 0000-0002-5708-727X.