World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Breast Cancer Classification using XGBoost

Rahmanul Hoque [1, *], Suman Das [2], Mahmudul Hoque [3] and Ehteshamul Haque [4]

[1] Department of Computer Science, North Dakota State University, Fargo, North Dakota, ND 58105, USA.
[2] Department of Mechanical Engineering, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh.
[2] School of Business, San Francisco Bay University, Fremont, CA 94539, USA.
[3] Department of Computer Science, Morgan State University, Baltimore, Maryland 21251, USA.
[4] Department of Computer Science, School of Engineering, San Francisco Bay University, Fremont, CA 94539, USA.

World Journal of Advanced Research and Reviews, 2024, 21(02), 1985–1994

Publication history: Received on 13 January 2024; revised on 22 February 2024; accepted on 24 February 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.2.0625

Abstract

Breast cancer continues to be one of the foremost illnesses that results in the deaths of numerous women each year. Among the female population, approximately 8% are diagnosed with Breast cancer (BC), following Lung Cancer. The alarming rise in fatality rates can be attributed to breast cancer being the second leading cause. Breast cancer manifests through genetic transformations, persistent pain, alterations in size, color (redness), and texture of the breast's skin. Pathologists rely on the classification of breast cancer to identify a specific and targeted prognosis, achieved through binary classification (normal/abnormal). Artificial intelligence (AI) has been employed to diagnose breast tumors swiftly and accurately at an early stage. This study employs the Extreme Gradient Boosting (XGBoost) machine learning technique for the detection and analysis of breast cancer. XGBoost provides an accuracy of 94.74% and recall of 95.24% on Wisconsin breast cancer Wisconsin (diagnostic) dataset.

Keywords: Training; Machine learning; Breast cancer; Data models; Classification algorithms; XGBoost; Feature Importance; Computer Aided Diagnosis; Artificial Intelligence

1. Introduction

Breast cancer is one of the primary causes of disease prevalence among both women and men. The presence of cancerous cells within the body is often indicated by early symptoms exhibited by the patient. The course of treatment and prognosis are contingent upon the specific type of malignant growth, the extent of its spread, and the patient's overall health condition [1-5]. Possible therapeutic approaches include surgical intervention, chemotherapy, and radiation therapy. The survival rate is determined by various factors such as the stage of cancer, general well-being, and other individual characteristics. Unfortunately, only a mere 14% of individuals diagnosed with breast cancer manage to survive for a period of five years. Indications that may raise suspicion of breast cancer encompass the presence of a lump, alterations in the skin, nipple discharge, swelling, breast pain, fatigue, insomnia, gastrointestinal difficulties, and respiratory distress [6-15].

The rapid increase in computational resources, driven by technological advancements, has led to the availability of high-dimensional data. This has made it possible to accurately diagnose and predict the prognosis and treatment outcome of breast cancers. Imaging techniques, such as mammography, ultrasonography, and magnetic resonance imaging, contribute to the characterization of malignant regions, providing valuable information for definitive diagnoses. Furthermore, molecular-based profiling data obtained through omics technologies, including transcriptomics, proteomics, and metabolomics, have played a role in identifying dysregulated molecular pathways in breast cancer. To

[*] Corresponding author: Ehsan Bazgir

fully utilize the clinical potential and gain insights into tumorigenesis and progression, sophisticated computational analyses are necessary alongside traditional statistical approaches.

Machine learning, a branch of artificial intelligence, has emerged as a computational analytical methodology to analyze the vast amount of data generated. Within machine learning, there are unsupervised and supervised methods. Unsupervised methods involve extracting features without prior knowledge of the expected results. Clustering techniques, for example, are commonly used to analyze omics data, such as subtype-specific expression patterns in transcriptome data. On the other hand, supervised methods mimic given results by combining observed features. Diverse machine learning techniques, for example, artificial neural networks, support vector machines, naïve Bayes algorithm, random forests, and decision trees, have been used for this purpose.

2. Literature Reviews

Recent studies have proven the ability of artificial intelligence to give accurate results that help specialists make better decisions due to its ability to capture details better than Humans. In [16], four of Machine Learning algorithms, which are Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Convolutional Neural Networks (CNN), with five different breast cancer datasets are tested and analyzed to verify their performance in a binary classification of breast cancer. The results show that CNN obtained higher accuracy than the other tested algorithms in this type of data. The accuracy obtained by RF, SVM, and KNN was 91%, 90% and 84%, respectively using Wisconsin diagnostic dataset in [16]. In [17], four machine learning models such as Linear regression, Random Forest, Multi-layer Perceptron and Decision Trees (DT) were applied. The performance in terms of accuracy, MLP was better as compared to other techniques in [17]. MLP technique also performed better than other techniques when Cross Validation metrics was used in breast cancer prediction [17].

According to Jin [18], KNN algorithm is one of the most frequently used classification algorithm in machine learning due to its simplicity and versatility in implementation.

Belciug et al. [19] presented a comparative study of cluster network, Self Organizing Map and K-means in the detection of breast cancer, using the Wisconsin Prognostic Breast Cancer (WPBC) dataset [20] in which k-means performed batter.

Chaurasia and Pal [21] examined the performance of artificial neural networks (ANNs), Logistic Regression (LR), and Dyadic decision trees (DDTs) in breast cancer recurrence prediction using the Breast Cancer Dataset.

Angeline [22] compared the performance of Naïve Bayes, Decision tree (C4.5), K-Nearest Neighbor and Support Vector Machine to find the preeminent classifier in Wisconsin Breast Cancer (WBC) to predict the primary site of cancer. As per the analysis, SVM performs better than other.

Abonyi [22] used fuzzy clustering algorithms in order to detect cancer on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Lavanya et. al. [24] uses a hybrid and dynamic approach to improve the classification accuracies of WDBC dataset to a relatively higher degree with 10 fold cross validation. As per the analysis [25], cross validation is most effective methodologies to estimate the performance of a ML model on a validation set.

Cruz [26] surveyed machine learning techniques and algorithms in cancer diagnosis and prognosis. Tan A.C. [27] performed bagged decision tree, C4.5 decision tree on micro array data of cancer and presented the performance analysis.

Tsirogiannis [28] applied bagging techniques using decision trees, neural networks and SVM on medical databases. As per the analysis, bagging techniques show better accuracy.

3. Methodology

In this section, the proposed methodology will be discussed in details.

3.1. Description of Dataset

The performance of the machine learning techniques is based on the availability of a suitable and valid dataset. The following dataset is being used in this research.

As per the documentation of the dataset, the dataset [20] encompasses attributes pertaining to the attributes of cellular nuclei. These attributes are derived from a digitized representation of a fine needle aspirate (FNA) of a breast mass. The dataset comprises 569 rows, where each row denotes a distinct digitized image of a breast mass, and 33 columns. Out of 569 rows, no column is missing data besides the "Unnamed: 32" column which only has null values. The "Unnamed: 32" column along with the "id" column has been removed since neither will be useful for our analysis. Each digitized image either contains benign (i.e. non-cancerous) or malignant (i.e. cancerous) breast cells. We should determine how many are benign and how many are malignant. The total number of malignant (M) and benign (B) samples is 212 and 357, respectively. So, out of 569 instances in our dataset: 357 contained benign breast cells (62.74%), 212 contained malignant breast cells (37.26%).

3.2. Visualization of Features

In order to visually represent the relationships between features for images of both benign and malignant cells, an effective approach would be to employ box plots. By grouping the data into sets of 10, we can gain insight into the distribution of the data and identify any outliers that may be present. It is worth noting that our dataset comprises 30 distinct features.

From figure 1, it is seen that the average radius of malignant breast cells in images is larger than benign cells. Other metrics related to cell size such as perimeter and area also share this relationship. However, there are outlier benign cells that just happen to have the larger average radius of a malignant cell. The mean fractal dimension box plots have very similar medians and therefore this feature probably lacks predictive power for diagnosis differentiation. The characteristics found within this particular subset, as depicted in figure 2, appear to be less distinct and therefore have less predictive power. It is possible that features such as Area_se, perimeter_se, and radius_se could hold some predictive value, although it is worth noting that there are still a considerable number of outlier benign cells that exhibit similar sizes to malignant cells. Furthermore, there is a small subset of malignant cells that possess a smaller size typically associated with benign cells, as evidenced by the lower whisker of the blue malignant box plot overlapping with the median value of the orange benign box plot. Figure 3 shows the worst feature values for malignant and benign breast cell images.

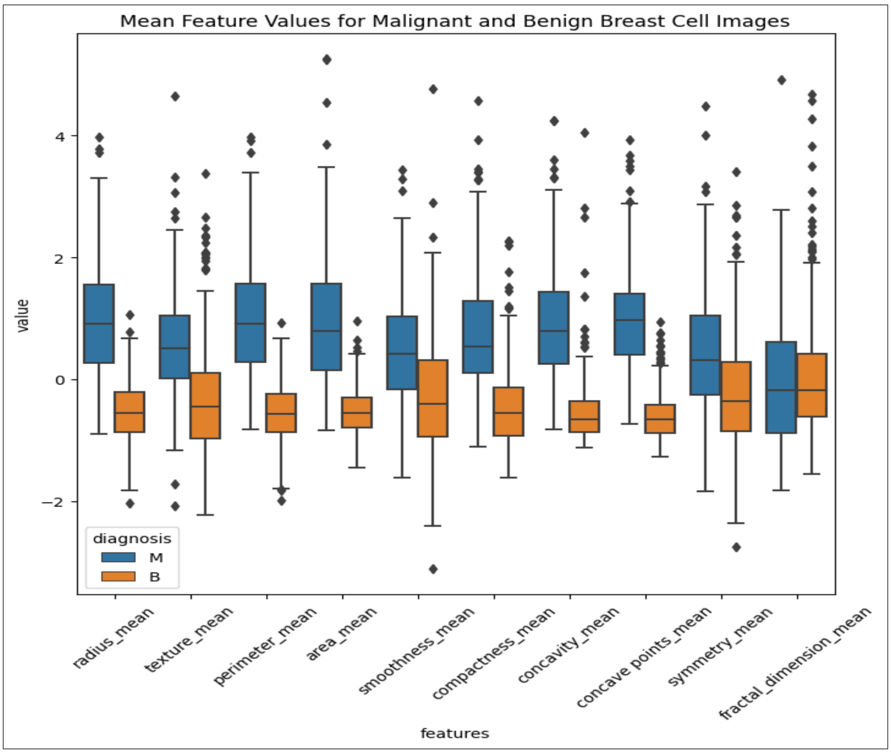

Figure 1 Mean feature values for malignant and benign breast cell

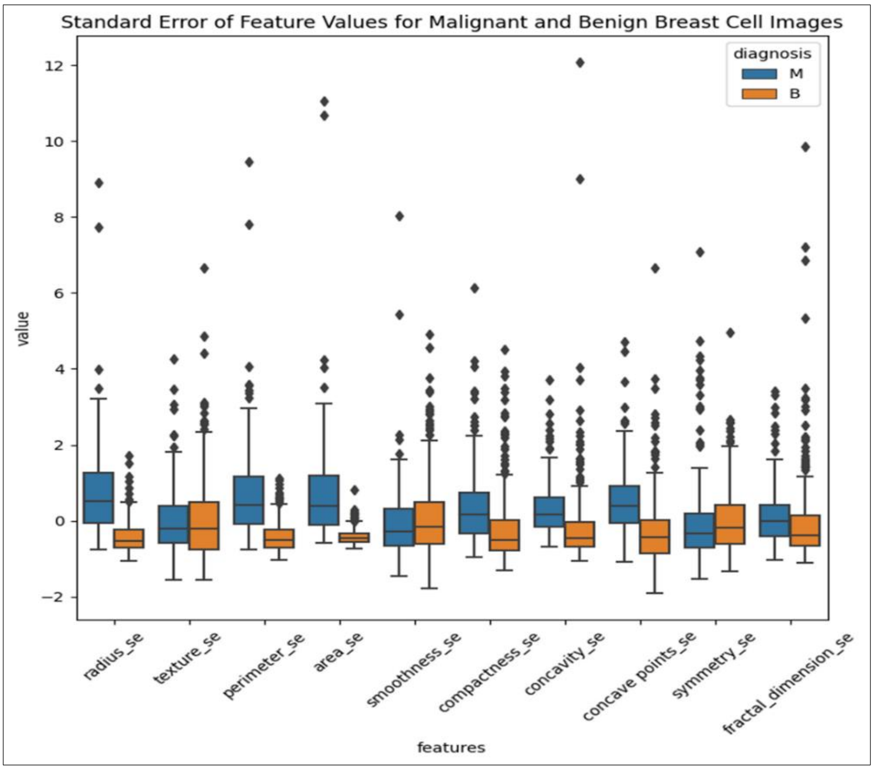

Figure 2 Standard error of feature values

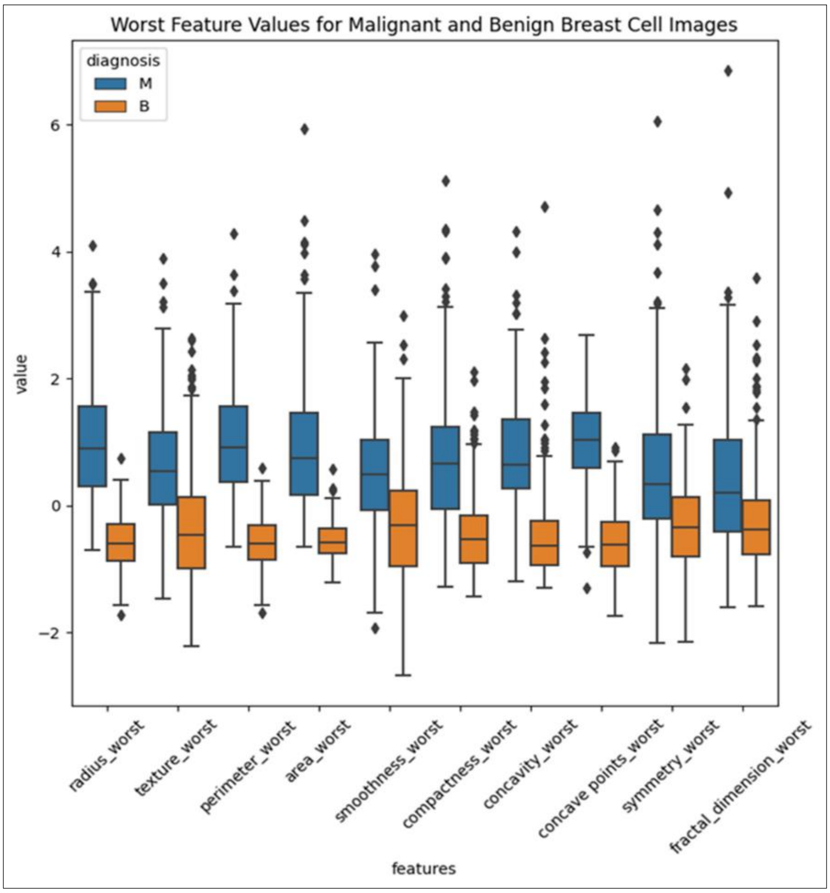

Figure 3 Worst feature values for malignant and benign breast cell images

The swarm plot shown in figure 4 provides a clearer representation of how outliers are blending in with the usual values for the other diagnosis. In the plot for fractal_dimension_mean, both malignant and benign observations are spread out across the entire plot. This relatively equal distribution once again suggests that this particular feature does not have much predictive ability. There is a more noticeable separation in the plots for radius_mean and area_mean. It is worth highlighting once again that features related to the size of the nucleus show potential predictive power. The plots for perimeter_mean and concave points_mean also appear to indicate some level of predictive ability.

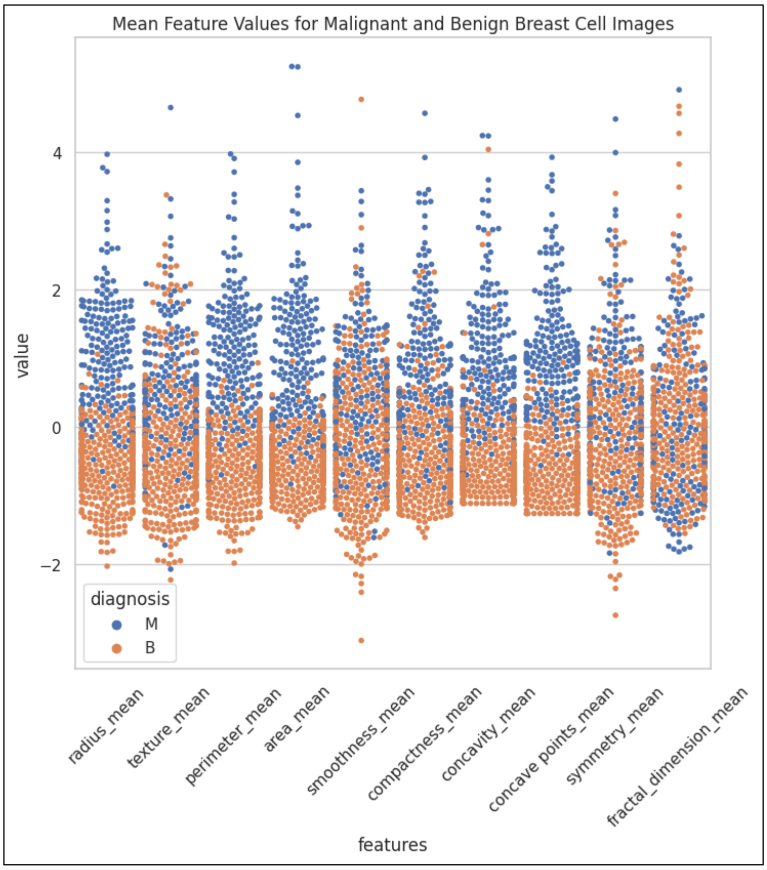

Figure 4 Swarm plots of mean feature values

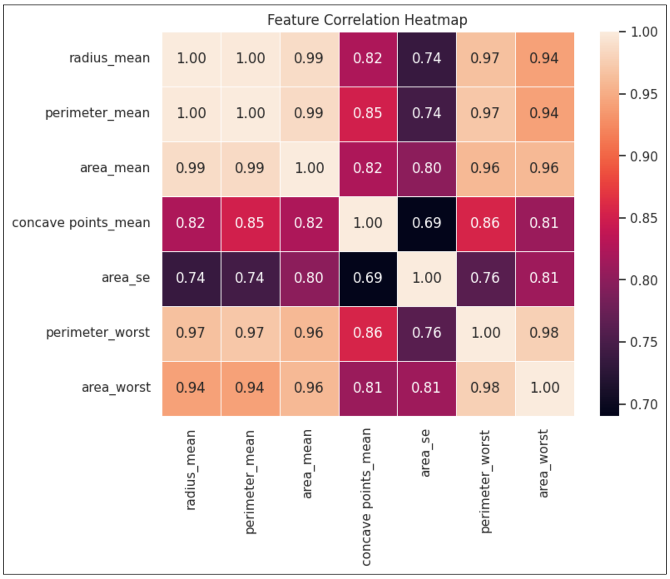

Figure 5 Feature Correlation Heatmap

Figure 5 describes the smaller correlation heatmap for the 7 features we identified as being most promising for predictive power.

3.3. Results of XGBoost

Many of the highly correlated features are very predictive. In contrast to linear regression models, machine learning models like XGBoost and Random Forest models are generally resistant to multicollinearity between features. Hence, for this problem, we will refrain from using a linear regression model. However, we choose XGBoost for the following reasons:

- With 30 features, XGBoost excels at managing the intricate relationships between them.
- XGBoost is capable of handling multicollinearity between features without compromising its performance.
- Compared to other machine learning options, XGBoost is known for its speed and efficiency due to its implementation of parallel processing.
- XGBoost provides a wider range of hyperparameters to fine-tune the model.
- Additionally, XGBoost offers the option to plot feature importance, which proves to be highly valuable for our research question.

The learning rate and max depth is chosen to 0.3 and 4. Testing dataset size is set to 20% and training dataset size is set to 80%. The model achieves an accuracy score of approximately 94.74%. Our model accurately predicts whether an image containing malignant or benign cells 94.74% of the time. The model's precision score is about 90.91%. It correctly identifies malignant cells in an image 90.91% of the time. The model's recall score is approximately 95.24%. It accurately identifies 95.24% of the images containing malignant cells in the testing data subset. The model's F1 score is about 93.02%. The F1 score represents the harmonic mean of the precision and recall scores.

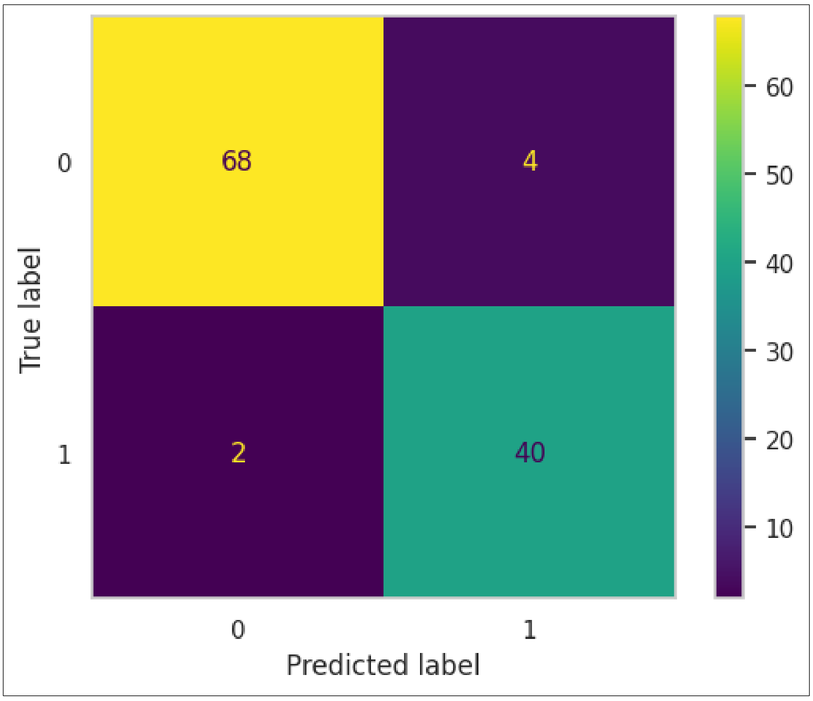

Figure 6 Confusion matrix for XGBoost model

Figure 6 shows the confusion matrix for XGBoost model. Out of the 114 unseen nuclei of breast cells in the subset for testing, our model achieves the following:

- 68 instances of True Negatives, where the model accurately predicts that an image contains benign breast cells.
- 2 instances of False Negatives, where the model inaccurately predicts that an image contains benign breast cells when they are actually malignant.
- 40 instances of True Positives, where the model accurately predicts that an image contains malignant breast cells.

- 4 instances of False Positives, where the model inaccurately predicts that an image contains malignant breast cells when they are actually benign.

3.4. Most Important Features to the XGBoost

After training the model to recognize patterns in breast cell nuclei, we found out which features are most important for telling apart benign and malignant cells. According to Figure 7, the model considers concave points_worst and area_worst as the most crucial, with both getting the highest F-score of 5.0 among the 14 features. Texture_worst and area_se also stand out with an F-score of 4.0 each. Gradient boosting refers to the process of "boosting" or strengthening one weak model by fusing it with a number of other weak models to create a more robust model as a whole. As an extension of boosting, gradient boosting formalizes the process of additively creating weak models as a gradient descent method over an objective function. To reduce errors, gradient boosting establishes desired results for the following model. The gradient of the error with regard to the prediction determines the desired objectives for each case, therefore the name "gradient boosting.

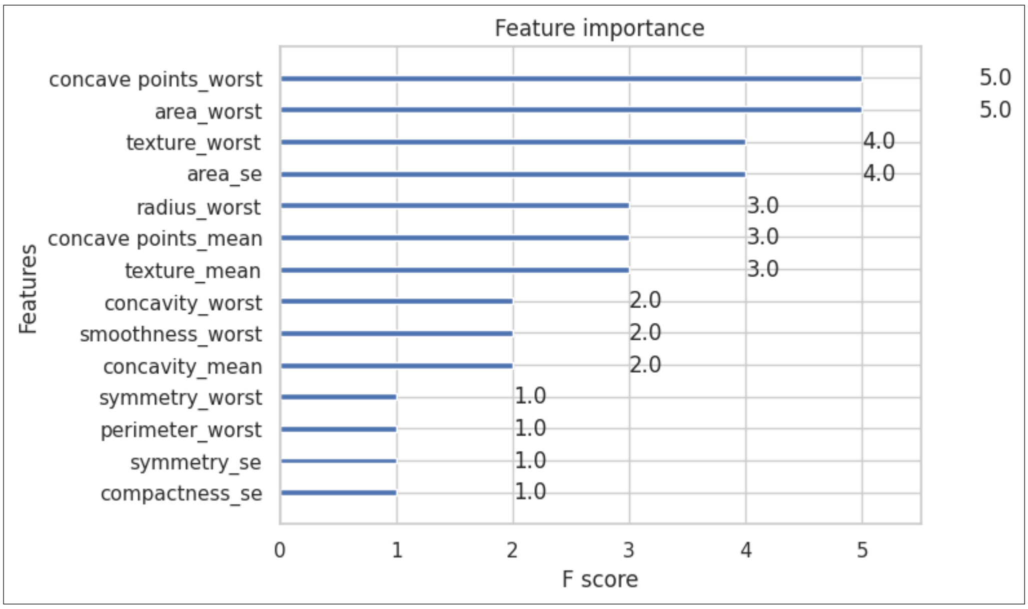

Figure 7 Feature ranking and F score

Table I shows the comparison between XGBoost and several existing systems based on accuracy as well as the number of instances and attributes were used. The table clarify that our proposed system with XGBoost algorithm is better than the existing systems in terms of performance.

Table 1 Comparison of proposed method with other models

Ref.	Sample	No. of features	Algorithm Name	Average Accuracy
[38]	256	5	SVM	83.3%
[39]	244	139	RF, K-stars, Neural Network	61.85%
[40]	569	32	Logistic Regression	94.4%
			Naive Bayes	92.3%
[41]	569	32	Decision Tree	94.4%
[42]	275	12	XGBoost	74%
			Random Forest	75%
Proposed XGBoost	569	32	XGBoost	94.74%

4. Conclusion

Most women suffer from breast cancer due to their unconsciousness. Various machine learning techniques are available to analyses medical data, but creating an effective and computationally efficient classifier is a significant machine learning challenge. We have used Extreme Gradient Boosting (XGBoost) ML algorithms in this paper. By applying this algorithm we have got accuracy of 94.74% accuracy for XGBoost. We will try to strengthen our work in future by handling a comparatively large dataset and incorporating some more functions such as breast cancer phase detection and so on. We hope that this study will contribute in the clinical application of breast cancer treatment.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] S. Bharati, M. A. Rahman and P. Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 581-584, doi: 10.1109/CEEICT.2018.8628084.

[2] Sifat Ibtisum, Ehsan Bazgir, S M Atikur Rahman, S. M. Saokat Hossain, "A comparative analysis of big data processing paradigms: Mapreduce vs. apache spark", World Journal of Advanced Research and Reviews, 20(01), pp. 1089-1098, 2023.

[3] Bipasha Sarker, Numair Bin Sharif, Mohammad Atikur Rahman, A.H.M Shahariar Parvez, "AI, IoMT and Blockchain in Healthcare", Journal of Trends in Computer Science and Smart Technology, Vol. 5, Issue:1, pp. 30-50, 2023.

[4] Rubaiyat Hossain Mondal, M., Subrato Bharati, and Prajoy Podder. "Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A review." arXiv e-prints (2021): arXiv-2110.

[5] Mondal, M. Rubaiyat Hossain, Subrato Bharati, and Prajoy Podder. "CO-IRv2: Optimized InceptionResNetV2 for COVID-19 detection from chest CT images." PloS one 16, no. 10 (2021): e0259179.

[6] Ahmmed, S., Podder, P., Mondal, M.R.H., Rahman, S.A., Kannan, S., Hasan, M.J., Rohan, A. and Prosvirin, A.E., 2023. Enhancing Brain Tumor Classification with Transfer Learning across Multiple Classes: An In-Depth Analysis. BioMedInformatics, 3(4), pp.1124-1144.

[7] Ehsan Bazgir, Ehteshamul Haque, Md. Maniruzzaman, Rahmanul Hoque, "Skin cancer classification using Inception Network", World Journal of Advanced Research and Reviews, 2024, 21(02), 839–849.

[8] Alam, F. B., Podder, P., & Mondal, M. R. H. (2023). RVCNet: A hybrid deep neural network framework for the diagnosis of lung diseases. Plos one, 18(12), e0293125.

[9] Rahman, S. M., Ibtisum, S., Bazgir, E., & Barai, T. (2023). The Significance of Machine Learning in Clinical Disease Diagnosis: A Review. arXiv preprint arXiv:2310.16978.

[10] Bazgir, E., Haque, E., Sharif, N. B., & Ahmed, M. F. (2023). Security aspects in IoT based cloud computing. World Journal of Advanced Research and Reviews, 20(3), 540-551.

[11] Ibtisum, S., Rahman, S. A., & Hossain, S. S. (2023). Comparative analysis of MapReduce and Apache Tez Performance in Multinode clusters with data compression. World Journal of Advanced Research and Reviews, 20(3), 519-526.

[12] Selim Molla, Ehsan Bazgir, S M Mustaquim, Iqtiar Md Siddique and Anamika Ahmed Siddique, "Uncovering COVID-19 conversations: Twitter insights and trends", World Journal of Advanced Research and Reviews, 2024, 21(01), 836–842.

[13] Mondal, M. R. H., Bharati, S., Podder, P., & Podder, P. (2020). Data analytics for novel coronavirus disease. informatics in medicine unlocked, 20, 100374.

[14] Mohammad Atikur Rahman, Ehsan Bazgir, S. M. Saokat Hossain and Md. Maniruzzaman, "Skin cancer classification using NASNet", International Journal of Science and Research Archive, 2024, 11(01), 775–785.

[15] Alam, F. B., Podder, P., & Mondal, M. R. H. (2023). RVCNet: A hybrid deep neural network framework for the diagnosis of lung diseases. Plos one, 18(12), e0293125.

[16] A. Bah and M. Davud, "Analysis of Breast Cancer Classification with Machine Learning based Algorithms," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, pp. 1-4, doi: 10.1109/ICMI55296.2022.9873696.

[17] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 997-1002, doi: 10.1109/ICCMC.2018.8487537.

[18] Marsilin, Jini R., and G. Wiselin Jiji. "An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier." Bonfring International Journal of Advances in Image Processing 2.1 (2012): 1.

[19] S. Belciug, F. Gorunescu, A. B. Salem, and M. Gorunescu., "Clustering-based approach for detecting breast cancer recurrence" In Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA), (2010):533–538.

[20] http://archive.ics.uci.edu/dataset/16/breast+cancer+wisconsin+prognostic (Accessed February, 2024)

[21] Chaurasia, Vikas, and Saurabh Pal. "Data mining techniques: To predict and resolve breast cancer survivability." International Journal of Computer Science and Mobile Computing 3.1 (2014): 10-22.

[22] Christobel, Angeline, and Y. Sivaprakasam. "An empirical comparison of data mining classification methods." International Journal of Computer Information Systems 3.2 (2011): 24-28.

[23] Abonyi, Janos, and Ferenc Szeifert. "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters 24.14 (2003): 2195-2207.

[24] Lavanya, D., and K. Usha Rani. "Ensemble decision tree classifier for breast cancer data." International Journal of Information Technology Convergence and Services 2.1 (2012): 17.

[25] http://www.openml.org/a/estimation-procedures/1 (Accessed February, 2024)

[26] Cruz, Joseph A., and David S. Wishart. "Applications of machine learning in cancer prediction and prognosis." Cancer informatics 2 (2006).

[27] Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification." (2003).

[28] Tsirogiannis, G. L., et al. "Classification of medical data with a robust multi-level combination scheme." Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. Vol. 3. IEEE, (2004).

[29] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (Accessed February, 2024)

[30] Rao, P. M., Singh, S. K., Khamparia, A., Bhushan, B., & Podder, P. (2022). Multi-class breast cancer classification using ensemble of pretrained models and transfer learning. Current Medical Imaging, 18(4), 409-416.

[31] Khamparia, A., Bharati, S., Podder, P., Gupta, D., Khanna, A., Phung, T. K., & Thanh, D. N. (2021). Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. Multidimensional systems and signal processing, 32, 747-765.

[32] Begum, A.M.; Mondal, M.R.H.; Podder, P.; Kamruzzaman, J. Weighted Rank Difference Ensemble: A New Form of Ensemble Feature Selection Method for Medical Datasets. BioMedInformatics 2024, 4, 477-488. https://doi.org/10.3390/biomedinformatics4010027

[33] Aggarwal, R., Podder, P., & Khamparia, A. (2022). Ecg classification and analysis for heart disease prediction using xai-driven machine learning algorithms. In Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI) (pp. 91-103). Singapore: Springer Singapore.

[34] Bharati, S., Podder, P., & Mondal, M. R. H. (2020, June). Diagnosis of polycystic ovary syndrome using machine learning algorithms. In 2020 IEEE region 10 symposium (TENSYMP) (pp. 1486-1489). IEEE.

[35] Podder, P., Mondal, M. R. H., & Kamruzzaman, J. (2022). Iris feature extraction using three-level Haar wavelet transform and modified local binary pattern. In Applications of Computational Intelligence in Multi-Disciplinary Research (pp. 1-15). Academic Press.

[36] Amit Deb Nath, Md Masum Billah, Numair Bin Sharif, Mahmudul Hoque, “Security Issues and Potential Solution of IoMT”, International Journal of Computer Applications, 2024.

[37] Bharati, S., Robel, M. R. A., Rahman, M. A., Podder, P., & Gandhi, N. (2021). Comparative performance exploration and prediction of fibrosis, malign lymph, metastases, normal lymphogram using machine learning method. In Innovations in Bio-Inspired Computing and Applications: Proceedings of the 10th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2019) held in Gunupur, Odisha, India during December 16-18, 2019 10 (pp. 66-77). Springer International Publishing.

[38] B. Majeed, H. T. Iqbal, U. Khan and M. A. Bin Altaf, "A Portable Thermogram based Non-contact Non-invasive Early Breast-Cancer Screening Device", 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), pp. 1-4, 2018.

[39] B. Bekta and S. Babur, "Machine learning based performance development for diagnosis of breast cancer", 2016 Medical Technologies National Congress (TIPTEKNO), pp. 1-4, 2016.

[40] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[41] T. Padhi and P. Kumar, "Breast Cancer Analysis Using WEKA", 2019 9th International Conference on Cloud Computing Data Science & Engineering (Confluence), pp. 229-232, 2019.

[42] S. Kabiraj et al., "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225451.