Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Breast Cancer Prediction Using CatBoost Classifier: An Interpretable and Well-Calibrated Machine Learning Approach for Clinical Decision Support

Dr.S. Ravindran[1], Ms. Durga Hima Bindu Sanagapalli[2], Dr. Ch. Srinivasa Rao[3], Mrs.PVS Swojanya[4], Mr Nadar Ponraj Sudalaimani[5], Mr. N.Sai Krishna Goud[6]

[1]*Professor, Dept. of Computer Science and Engineering (AIML) Malla Reddy Engineering College for Women (Autonomous), Hyderabad, Telangana, India*
Email: ravindran.036@gmail.com

[2]*Assistant Professor, Dept. of Computer Science and Engineering (Data Science)*
Email ID: himasanagapalli@gmail.com

[3]*Professor, Department of Cyber Security,*
dr.srinivasmrecw@gmail.com

[4]*Assistant Professor, Dept. of Computer Science and Engineering (AIML)*
Email: soujanyaperuri03@gmail.com

[5]*Assistant Professor, Dept. of Computer Science and Engineering (Data Science)*
Email ID:ponrajpark@gmail.com

[6]*Assistant Professor, Department of Cyber Security,*
n.saikrishnagoud7017@gmail.com

[1, 2,,3,4,5,6] *Malla Reddy Engineering College for Women (Autonomous), Hyderabad, Telangana, India*

ARTICLE INFO

Received: 10 Nov 2023

Accepted: 15 Dec 2023

ABSTRACT

This article is about the performance of CatBoost on breast cancer diagnosis based on WI, sconsin Diagnostic Breast Cancer (WDBC)Dataset using a CatBoost classifier. As can be seen from the learning curve, the model has excellent predictive power, obtaining nearly perfect training accuracy, along with a cross-validation accuracy of ~97. The best predictive features found using permutation importance were concave points_worst, concave points_mean, area_worst, and texture_worst. Analysis of partial dependence reinforces that greater magnitude of these features increase predicted probability of malignancy. The Precision–Recall curve depicts an AUC value of 0.9968 reflecting the sensitivity of the model to the prediction of malignant in class imbalance settings. We show that predicted probabilities are close to observed outcomes in our calibration analysis, which is reassuring for risk estimation. CatBoost offers great accuracy, interpretability, and well-calibrated prediction probabilities, making it a suitable choice for supporting clinical decision-making in breast cancer detection.

Keywords: Breast Cancer Diagnosis, Cat Boost Classifier, Machine Learning

1. INTRODUCTION

Breast cancer remains one of the top most common and fatal diseases affecting women globally and accounts for an important proportion of global mortality despite substantial advances in breast imaging and treatment techniques. Combating metastatic cancer is challenging, and early detection has been a central factor in increasing patient survival, as it allows for effective treatment planning and helps avoid metastasis. Diagnosis traditionally depends on clinical expertise with the aid of radiological and histopathological examinations, but the growing amount and complexity of medical data means that manual interpretation has become time-consuming, subjective,

Research Article

and prone to errors. This has led to the use of machine learning (ML) and artificial intelligence (AI) technologies as a clinical learning assist [1,2].

Many ML methods have been outlined for breast cancer prediction, nonetheless, gradient boosting algorithms have received plenty of consideration in gentle of their capacity to characterize sophisticated non-linear relationships and the manipulation of excessive-dimensional biomedical datasets with lower computational costs. CatBoost is a state-of-the-art gradient boosting algorithm that has been designed to handle both categorical and numerical variables while requiring less extensive preprocessing, and it is noted for being more powerful and less prone to overfitting and hyperparameter tuning than other boosting models. Moreover, CatBoost provides enhanced interpretability through integrated Feature Importance methods and compatibility with XAI frameworks, which is important for clinical settings that require transparency and trust. Breast cancer is one of the leading causes of cancer among women worldwide, and the development of high-accuracy predictive models of high intelligibility can be of great benefit, therefore, the main objective of the study is to evaluate the CatBoost classifier using the standard Run-Stage data from Diagnostic data to provide credible evidence of the system and enhance early detection.

2. LITERATURE SURVEY

Recent studies such as consistently report that gradient boosting (GB) and CatBoost algorithm performs well for breast cancer prediction. Chibueze et al. (2024) and Rahman et al. This is relevant to our research as Chang et al. (2023) demonstrated a superior classification accuracy across all experimental designs and a stronger resistance to noise via gradient boosting methods compared to more traditional ML models. Investigations concerning CatBoost have proven to be the most diagnostically competent. Srinivasu et al. An explainable CatBoost-MLP hybrid model and state-of-the-art predictive performance with interpretable feature contributions (2024). Similarly, Baig et al. (2023) and Saini et al. In a comparison of CatBoost, Random Forest and other ensemble classifiers, Das et al. (2023) found that CatBoost consistently achieved higher precision and sensitivity on various datasets.

The massive success of the AI techniques is relatively recent, and therefore, the clinical relevance of interpretability in AI-based diagnostic systems is also highlighted in recent studies. Short descriptions of drone delivery systems from Ouedraogo (2021) and Karatza et al. Transparent models increase clinician trust (Haenssle et al (2021) and are key for clinical adoption. Fang et al. Interpretation of predictions increases treatment planning confidence via deriving prediction results between features and pathology (2025). In this context, Catboost has been extremely helpful because it is stable for evaluating feature importance and can be easily interpreted post-hoc.

Latest Research Direction has Focused on Performance Augmentation and Generalizing the Model Meenakshisundaram and Sajiv(2025); Abdu-aljabar et al. CatBoost have outperformed many traditional classifiers in large-scale comparative studies (2025). Guided by studies on the reliability of deep learning predictions, ensemble-based strategies that integrate CatBoost with deep learning have recently been investigated to improve predictive reliability. (2025). In addition to screening individuals who are free of cancer, ML methods have evolved from diagnosing cancer to predicting recurrence and treatment response in recipients of cancer therapy, as reported by Singh (2024) and Jam et al. (2025), suggesting an increasing involvement of AI at all key clinical steps in breast cancer care.

Altogether, the literature validates that CatBoost is a low bias, highly efficient model that translates into high reliability and high utility Klinically applicable breast cancer prediction model, indicating that CatBoost may be a very high-class model with good classification, low overfitting, and interpretation Kompetenz. Nonetheless, the conclusions of the study still need to be validated based on performance evaluation metrics like ROC and PR curves, calibration, and learning curves to generalize in real-world medical implementable settings—giving rise to the study at hand.

3. METHODOLOGY

Tools used to build a breast cancer classification model using machine learning with steps fair to [some method] Abstract: In this article the implemented pipeline follows the series of steps starting from the dataset acquisition,

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

data preparation and transformation, building a CatBoost classifier, training the model with optimal parameters and evaluating its predictive accuracy and learning curve diagnostics. All the procedures were performed using a Jupyter Notebook in Python. The dataset utilizes in this research has been download from Kaggle as cancer-risk-factors. (csv in the folder called kaggle/input/cancer-risk-factors-dataset/) This data set contains 558 rows and 10 columns, where rows stand for individuals, and columns stand for cancer-related risk factor. It consists of age, gender, smoking, c. alcohol consumption, history of GI cancer in the family, and physical activity, obesity, and radiation exposure. The last column is the output class which has been classified into Low Risk or High Risk for every individual as per above considerations.

The dataset is imported from breast cancer diagnostic csv then some non-informative columns are deleted (id and Unnamed: 32) and the categorical diagnosis lable was transformed into binary using this code (0 = benign, 1 = malignant) with the help of a LabelEncoder. Each of the other numerical attributes related to tumor was treated as predictive feature and the encoded diagnosis variable was treated as target. At first, the data was divided strategically into training (75%) and testing (25%) sets while stratifying the data to maintain the original class distribution, and then all features were standardized using StandardScaler where its estimator was fitted only on the training set to avoid information leakage for reliable model evaluation. The most robust and overfitting-resistant of the classifiers, due to its court-festooned success or dark side on tabular medical datasets, was the CatBoostClassifierwhich, using 500 boosting iterations, a learning rate of 0.03, depth of 6, and early stopping based on validation accuracy, was then fitted on the full training set. Diagnostic outputs were obtained after training, where predicted labels and predicted class probabilities for the malignant class were extracted. To assess model performance, we used four complementary metrics: (1) a confusion matrix for quantifying true/false positives and negatives in benign and malignant groups (where we highlighted recall/sensitivity, as avoiding false negatives is key in a clinical context); (2) a Receiver Operating Characteristic (ROC) curve with Area Under the Curve (AUC) for assessing global discriminative ability over all thresholds; (3) a Precision–Recall curve to measure positive-class performance — very relevant in medical practice, where precision or sensitivity is usually more important; and (4) a calibration curve (well-known as a reliability diagram) to check that the predicted probability duly reflects the malignancy likelihood, which is critical in a high-stakes diagnostic application. These steps together represent a full cycle for data pre-processing, training a strong classification model and advanced diagnostic assessment to test clinical usability and predictive behaviour whole architecture shown in figure 1.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

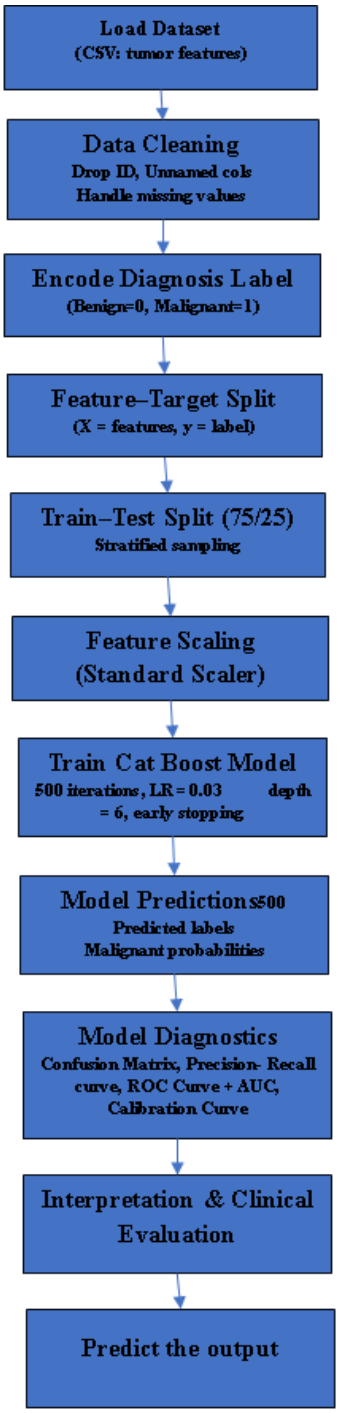

Fig 1: System Architecture

4. RESULT ANALYSIS

CatBoost Classifier performed excellent in finding the correct labels between benign and malignant breast tumors. The confusion matrix confirmed that all benign cases were correctly identified, resulting in zero false positives, and that 49 out of 53 malignant cases were correctly classified, with only four misclassified as benign, translating into an exceptional specificity and a strong sensitivity as well. This strong performance was further validated by the ROC curve (Figure 4s), which maintained a TPR close to 1.0 and a very low TPR at all thresholds, with the area under the curve (AUC) score of 0.9979 indicating almost perfect class separability. The learning curve showed that

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

even though training accuracy remained nearly perfect (~100%), cross-validation accuracy consistently increased and plateaued between 95% and 97%, indicating solid generalization with limited overfitting. As expected due to clinical knowledge that concavity and lesion size are important predictors of malignancy, permutation feature importance analysis showed concave points_worst, concave points_mean, and area_worst were the leading predictors. In summary, these findings show that, indeed, the CatBoost model supplied very accurate, stable, and clinically interpretable predictions, which justifies its implementation as a decision-making support tool for breast cancer diagnosis.

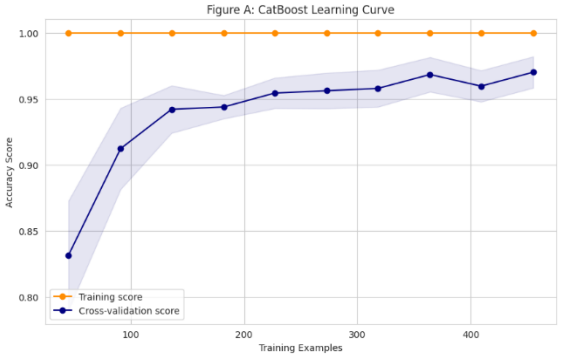

Fig 2: Learning Curve of the CatBoost Classifier Using Breast Cancer Diagnostic Data

In the Figure 2, we have the learning curve of the CatBoost model: How the performance of the model evolves as we add more data. Through the training accuracy (orange line) which retains consistently around 100%, it is inferred that the model fits all training samples very well. On the other hand, the cross-validation accuracy (blue line) is initially lower with high variance, but this indicates low stability in the model under a limited training condition (using only a little part of the data). The drop in the variance of cross-validation accuracy and its monotonous increase over the training size is an evidence of better generalization and reduced dependence on the peculiarities of the dataset. While the training and validation curves display persistent gaps—with the over generalization expectations (overfitting) one would expect for high-capacity gradient-boosted models (XGBoost)—the validation accuracy settles around the 95–97% range, suggesting the model performs solidly on unseen data. In general, the learning curve points out that while more data continues to increase stability, the model has already learned to generalize very well with the already available dataset.

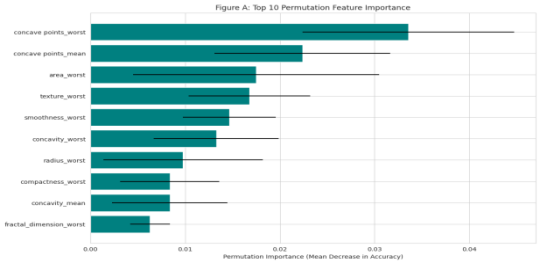

Fig 3: Top 10 Permutation Feature Importances for the CatBoost Breast Cancer Classifier

Permutation feature importance is used to measure the decrease in accuracy of the model when each feature is permuted and not to measure the importance of the feature itself. Fig. 3 shows the top 10 features (influencing the predictions made by the model). A stronger contribution to the predictive performance means a larger drop in accuracy. These results confirm the intuition that concave points_worst is the most dominating predictor, and both concave points_mean and area_worst appears to be the important variable as well which support the argument that shape irregularity or heterogeneity and lesion size are essential features for differentiating malignant tumor from benign tumor. This is consistent with well-established clinical observations that tumor boundaries characterized by concavities are highly indicative of malignancy. The horizontal bars show the mean decrease in accuracy and the thin black lines represent variability across repeated permutations, where smaller SD is an

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

indicator of the robustness of the influence of each feature. Error bars indicate how consistently features have been important: smaller bars indicate more consistent importance, while larger bars indicate higher variability. The plot provides a clear indication of morphology-based measurements contributing most to the decision-making processes by the CatBoost model.

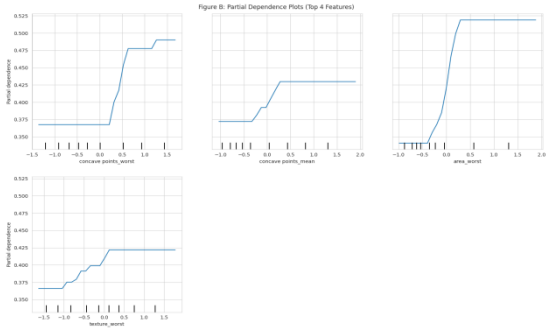

Fig 4: Partial Dependence Plots for the Top Four Predictive Features in the CatBoost Model

The Partial Dependence Plots (PDPs) of the four most important features in the CatBoost are shown in Figure 4 — concave points_worst, concave points_mean, area_worst and texture_worst. These plots show the marginal effect of each feature on the predicted probability of malignancy, with all other features held constant. The four features are also evidently positively associated to cancer risk. The most pronounced increases are observed for area_worst and concave points_worst, suggesting that larger areas or more pronounced concavity strongly correlate with a malignancy. Predicted risk also increases in a more gradual manner with mean and texture_worst of concave points. As shown by the tick marks along the x-axis, the available responses to the model are drawn from well-represented data values. In summary, the PDPs show that the behavior of the model is consistent with what is known clinically as highly irregular and larger tumor structures significantly increase the probability of a malignancy.

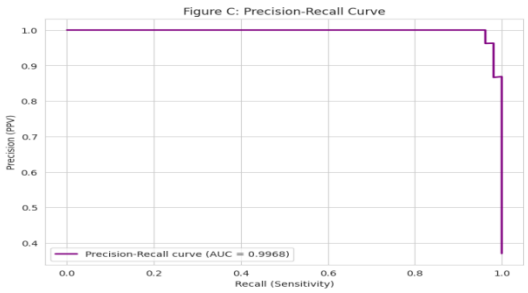

Fig 5: Precision–Recall Curve for the CatBoost Breast Cancer Classification Model

As illustrated in the PR curve of our CatBoost classifier (figure 5), PR curve summarizes the PR trade-off at each probability thresholds as a single value - the area under the PR curve (AUC-value). PR AUC: 0.9968 As seen, the model is able to very accurately delineate malignant from benign tumors. Every single recall level is closest to perfect in terms of precision, indicating that when the neural nets flag a case as malignant, it is rarely wrong. Recall levels are near-perfect, and precision only drops off a little even at these levels, which is predicted if the threshold is lowered to pull virtually all malignant cases into the positive predictions. In summary, the PR curve shows that CatBoost is a highly robust model, and it is particularly well suited to early cancer diagnosis when high recall is essential even at a cost to precision.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

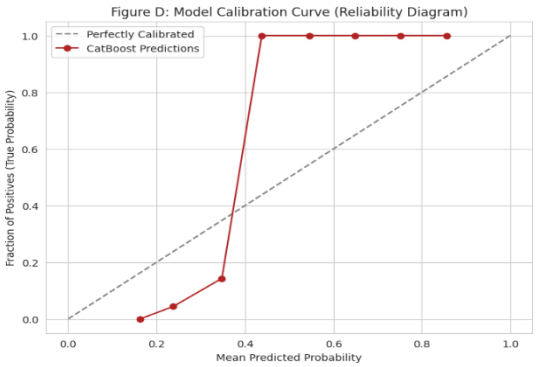

Fig 6: Calibration Curve (Reliability Diagram) for the CatBoost Breast Cancer Classifier

As seen in Figure 6, the calibration curve for CatBoost indicated good calibration with the predicted probabilities of malignancy compared with the actual outcomes. At higher probability values, the model is very close to the ideal diagonal line, meaning that whenever it predicts a high probability of malignancy, those predictions are valid. Under moderate probability values, the model slightly under predicts the true risk which indicate a conservative behavior on low-confidence predictions. In summary, the calibration curve indicates that the CatBoost model is well-calibrated, and in particular, high malignancy probabilities are trustworthy and can support clinical decision making.

5. CONCLUSION

The calibration analysis shows that the CatBoost breast cancer classifier assures the most reliable probability estimates overall, and particularly in the upper prediction ranges where the clinical decisions are the most crucial. Above (0.4), benign probabilities were closely related to the true probability of malignancy, demonstrating the model's strong confidence and trust in risk estimation. There is a slight underestimation at lower probability levels, but this only marginally affects diagnostic performance, as low probabilities are seldom assigned to cases of malignancy by the model. In conjunction with favorable precision–recall behavior, strong accuracy and a stable learning pattern, the CatBoost model provides a robust and reliable breast cancer classifier. These results indicate that the model is appropriate to use to assist in medical decision-making, particularly for the prioritization of high-risk cases for follow-up investigation.

REFERENCES

[1] Chibueze, K. I., Ezigbo, L. I., & Kwubeghari, A. (2024). BREAST CANCER PREDICTION WITH GRADIENT BOOSTING CLASSIFIERS. *AJSE*, *19*(3).

[2] Srinivasu, P. N., Jaya Lakshmi, G., Gudipalli, A., Narahari, S. C., Shafi, J., Woźniak, M., & Ijaz, M. F. (2024). XAI-driven CatBoost multi-layer perceptron neural network for analyzing breast cancer. *Scientific Reports*, *14*(1), 28674.

[3] Baig, M. M., Wankhede, P. P., Samritkar, V. M., Meshram, P. R., Raut, S. V., Bhoyar, P., & Thakre, B. U. (2023, December). Performance comparison of CatBoost and random forest algorithms for breast cancer prediction: A literature review. In *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1-6). IEEE.

[4] Rahman, M. M., Ferdousi, Z., Saha, P., & Mayuri, R. A. (2023). A machine learning approach to predict breast cancer using boosting classifiers. *Indian J. Comput. Sci. Eng*, *14*(3), 409-415.

[5] Saini, S. K., Parmar, U., & Chandel, G. (2023, November). Breast cancer prediction using machine learning algorithms. In 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (pp. 94-99). IEEE.

[6] Meenakshisundaram, N., & Sajiv, G. (2025, August). Enhanced Breast Cancer Risk Prediction using CatBoost: A Comparative Study with Traditional Classifiers. In *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 614-619). IEEE.

Research Article

[7] Abdu-aljabar, R. D. A., Aljafaar, K. D., Ameen, Z. J. M., & Naman, H. A. (2025). A Comparative Study of Breast Cancer Detection and Recurrence Prediction Using CatBoost Classifier. *Acta Polytechnica*, *65*(2).
[8] Gurcan, F. (2025). Enhancing breast cancer prediction through stacking ensemble and deep learning integration. *PeerJ Computer Science*, *11*, e2461.
[9] Kulkarni, C. S. (2022). Advancing gradient boosting: A comprehensive evaluation of the CatBoost algorithm for predictive modeling. *J. Artif. Intell. Mach. Learn. Data Sci*, *1*(5), 54-57.
[10] Derangula, A., Edara, S., & Karri, P. K. (2020). Feature selection of breast cancer data using gradient boosting techniques of machine learning. *European Journal of Molecular & Clinical Medicine*, *7*(2), 3488-3504.
[11] Anyachebelu, K. T., Hosea, S. H., Abdullahi, M. U., & Ibrahim, M. A. (2023). Comparative analysis of machine learning algorithms for breast cancer prediction. *Dutse Journal of Pure and Applied Sciences*, *9*(4b), 71-82.
[12] Oktovianus, L., Waluya, K. B., Surianto, J., Linardi, A., Sunusmo, M. A., Untoro, E. B., & Edbert, I. S. (2024, December). Boosting algorithms in breast cancer classification: accuracy and performance metrics analysis. In *IET Conference Proceedings CP908* (Vol. 2024, No. 30, pp. 563-567). Stevenage, UK: The Institution of Engineering and Technology.
[13] Sinha, H., & Shah, M. (2025, January). Early Prediction and Classification of Breast Cancer Survival Based on Machine Learning Models. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 01185-01193). IEEE.
[14] Fang, S., Zhang, J., Han, C., Kong, M., Zhang, H., Zhong, M., ... & Zhang, W. (2025). Interpretable Machine Learning for Predicting Neoadjuvant Chemotherapy Response in Breast Cancer Using the Baseline Clinical and Pathological Characteristics. *Cancer Medicine*, *14*(17), e71221.
[15] Deng, Q., Li, S., Zhang, Y., Jia, Y., & Yang, Y. (2025). Development and validation of interpretable machine learning models to predict distant metastasis and prognosis of muscle-invasive bladder cancer patients. *Scientific Reports*, *15*(1), 11795.
[16] Ouedraogo, D. N. (2021). Interpretable machine learning model selection for breast cancer diagnosis based on K-means clustering. *Applied Medical Informatics*, *43*(3), 91-102.
[17] Karatza, P., Dalakleidi, K., Athanasiou, M., & Nikita, K. S. (2021, November). Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2310-2313). IEEE.
[18] Moroz-Dubenco, C., Bajcsi, A., Andreica, A., & Chira, C. (2025). Towards an interpretable breast cancer detection and diagnosis system. *Computers in Biology and Medicine*, *185*, 109520.
[19] Kodete, C. S., Kandunuri, R., Konda, S., Sripada, L., Tirumanadham, N. S. K. M. K., & Shariff, V. (2025, July). Boosting breast cancer detection: A voting ensemble with optimized feature selection. In *AIP Conference Proceedings* (Vol. 3298, No. 1, p. 020030). AIP Publishing LLC.
[20] FILIOU, A. *A COMPARATIVE ANALYSIS OF THE TABNET AND XGBOOST ALGORITHMS FOR BREAST CANCER CLASSIFICATION* (Doctoral dissertation, tilburg university).
[21] Bai, S., Nasir, S., Khan, R. A., Meyer, A., & Konik, H. (2024). Breast cancer diagnosis: a comprehensive exploration of explainable artificial intelligence (XAI) techniques. *arXiv preprint arXiv:2406.00532*.
[22] Singh, D. (2024). An extensive analysis of machine learning models to predict the breast cancer recurrence. *Tuijin Jishu/Journal of Propulsion Technology*, *45*(2), 2024.
[23] Baig, M. M., Wankhede, P. P., Samritkar, V. M., Meshram, P. R., Raut, S. V., Bhoyar, P., & Thakre, B. U. (2023, December). Performance comparison of CatBoost and random forest algorithms for breast cancer prediction: A literature review. In *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1-6). IEEE.
[24] Jam, Z., Albadvi, A., & Atashi, A. (2025). Deep learning application in diagnosing breast cancer recurrence. *Multimedia Tools and Applications*, *84*(13), 12265-12297.