Predicting Breast Cancer by Integrating Clinical and MRI Data via Machine Learning

Aditya Purswani[1], Anasuya Dutta[1], Anurag Phukan[1], Dhruv Bhattarcharjee[1], and Shashwat Sinha[1],
[1]School of Computer Science, University of Nottingham

Using clinical and MRI data from the I-SPY 2 TRIAL, this study uses machine learning to determine Pathological Complete Response (PCR) and also (RFS) which is used to denote Relapse-Free Survival in patients with breast cancer. preprocessing the data, normalization, and feature optimization approaches such as Recursive Feature Elimination were used in the study. For PCR prediction, models such as Random Forest, SVM, Logistic Regression, KNN, etc were utilized, along with SMOTE for class balance. Random Forest Regressor, Gradient Boosting Regressor, etc were used to forecast RFS, with an emphasis on feature importance. Both models were assessed using balanced classification accuracy and Mean Absolute Error, revealing the promise of machine learning in personalized cancer therapy and decision-making.

Index Terms—**Breast Cancer Prognosis, Pathological Complete Response (PCR), Relapse-Free Survival (RFS)**

I. Introduction

Breast cancer, the most prevalent cancer in women globally, presents substantial challenges in terms of treatment and prognosis. Accurately determining the effectiveness of chemotherapy in achieving Pathological Complete Response or (PCR) and predicting Relapse-Free Survival or (RFS) is crucial for tailoring treatment strategies and improving patient outcomes. Recent advancements in machine learning offer promising avenues to enhance prediction accuracy by analyzing clinical and MRI-based data from the I-SPY 2 TRIAL dataset. However, issues like missing data, high dimensionality, and class imbalance necessitate extensive data preparation and feature engineering. This research employs various machine learning models to address these challenges, utilizing techniques like Synthetic Minority Over-sampling (SMOTE) and Recursive Feature Elimination to optimize classifier and regressor performance. The study also investigates the relevance of features in model predictions, with the ultimate goal of contributing to personalized medicine and informed clinical decisions in breast cancer therapy by achieving balanced classification accuracy for PCR and minimizing Mean Absolute Error for RFS.

II. Literature Review

The review of the literature covers various approaches of machine learning (ML) techniques for breast cancer detection and classification. A comprehensive method involving mammography screening images, morphological operations for tumor segmentation, and Random Forest (RF) classification with a 95% accuracy [1]. Dataset of Wisconsin Diagnostic of Breast cancer was used, with SVM and Artificial Neural Network demonstrating superior performance, with an impressive accuracy of 98.08% [2]. When all ML methods for breast cancer diagnosis are compared, SVM achieves the best accuracy among the evaluated techniques [3].

Using an Artificial Neural Network and a logistic algorithm, an adaptive ensemble voting method for breast cancer diagnosis achieved 98.50% accuracy [4]. Convolutional Neural Networks outperform other tested algorithms in accuracy when four ML algorithms are tested on five breast cancer datasets [5]. The use of multiple ML algorithms to classify as benign tumours or malignant type emphasizes the importance of early detection [6]. On the dataset of Wisconsin Diagnostic Breast Cancer dataset, a comparison of five ML algorithms, including Decision Tree, Support Vector Machine, Logistic Regression, Random Forest, and K-Nearest Neighbor, is performed [7].

The importance of ML in predicting breast cancer is highlighted, with Support Vector Machine achieving 96.25% accuracy [8]. The classification capabilities of logistic regression, decision trees, random forest, and CNN for breast cancer are investigated and compared [9]. A thorough comparison of seven ML classification techniques reveals XGboost to be the best performer, with an accuracy of 98.24% [10]. All of these studies demonstrate the evolving landscape of machine learning (ML) applications in breast cancer detection, with various algorithms demonstrating promising results in terms of accuracy and performance metrics.

III. Proposed Methodology

Various algorithms have been applied during the process. Bagging Classifier was demonstrated to be the most successful model for PCR prediction in breast cancer patients, with a high balanced accuracy of 87.25%. This was closely followed by the Random Forest and AdaBoost Classifiers, both of which attained an 80.39% balanced accuracy. Gradient Boosting achieved 73.53% accuracy, whereas K-Nearest Neighbours (KNN) achieved a balanced accuracy of 82.35%. The SVM and Logistic Regression models both attained 66.18%, indicating potential for improvement.

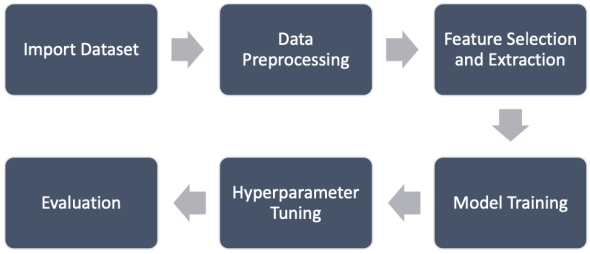


Fig. 1: Methodology

Notably, the Artificial Neural Network (ANN) model performed poorly in this setting, with a balanced accuracy of 50%. Gradient Boosting Regressor produced a Mean Absolute Error (MAE) of 21.47 in the regression challenge for predicting Relapse-Free Survival (RFS). These results tell us the efficacy and potential of several ML algorithms in personalized cancer therapy.

A. Dataset

The dataset, derived from the I-SPY 2 TRIAL, is a comprehensive collection designed to further research into breast cancer prognosis. It has 117 different characteristics for each patient, separated between 10 clinical and 107 MRI-based qualities. These parameters include age, hormone receptor status (ER, PgR, HER2), Triple-Negative status, chemotherapy grade, tumor proliferation rates, histological type, lymph node status, and tumor stage, as well as a plethora of radiomic data collected from MRI images. The incorporation of radiomic data is notable because it provides a precise and quantified study of tumor features, perhaps revealing trends not detectable from traditional clinical data alone. Furthermore, the dataset tackles a major difficulty in breast cancer treatment: predicting Pathological Complete Response PCR and RFS. This dataset's heterogeneous character, including clinical insights and sophisticated imaging data, provides a solid platform for employing machine learning approaches to predict treatment results and aid in personalized therapy decisions.

B. Data Preprocessing

In preparing the I-SPY 2 TRIAL dataset for the analysis, several crucial preprocessing steps have been undertaken. Firstly, missing values marked as '999' were transformed to NaNs and imputed with the median value of each feature, a strategy known for its robustness in datasets with potential outliers or skewed distributions. Outliers were not removed due to the data being sensitive and it could lead to loss of important information. The range of features were standardized, including both clinical and MRI-based data, through standard scaling, normalizing the data to mean zero and one as standard deviation.

This was especially vital for scale-sensitive models. Additionally, the high dimensionality of the dataset necessitated feature selection, for which Recursive Feature Elimination (RFE) has been used to distill the most impactful features. To tackle class imbalance in PCR prediction, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized, equalizing the representation of different classes. Finally, the data has been split into distinct training and testing sets to ensure both robust training and reliable evaluation of the models. These preprocessing steps were fundamental in rendering the dataset optimal for the subsequent machine-learning endeavors.

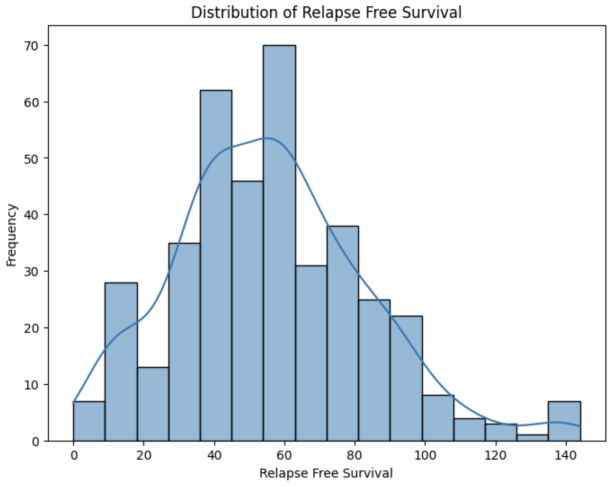


Fig. 2: Distribution Plot of RFS Outcome

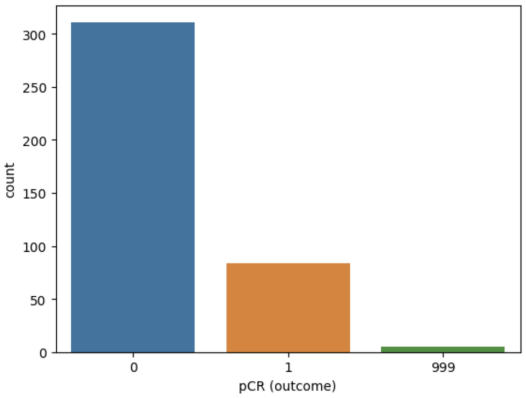


Fig. 3: Distribution Plot of PCR Outcome

C. Model Architect

This study leverages the I-SPY 2 TRIAL dataset for breast cancer prognosis, diverse array of ML models is employed to calculate PCR and RFS. The classifiers included Random Forest, SVM, Logistic Regression, and K-Nearest Neighbors (KNN), each chosen for their specific strengths in handling high-dimensional data and efficiency in classification. The KNN model was further enhanced with a Bagging Classifier to improve accuracy and reduce overfitting. For PCR prediction, AdaBoost with Decision Trees has been incorporated leveraging its boosting capabilities. The regressor models for RFS prediction featured Random Forest and Gradient Boosting, known for their robustness in regression tasks, along with SVR for its aptness in continuous output prediction. Additionally, AdaBoost and Bagging techniques were integrated with

Random Forest and Decision Trees, respectively, to augment their predictive performance. This comprehensive approach, evaluating models based on balanced classification accuracy and Mean Absolute Error, was meticulously crafted to ensure precise and reliable outcomes, aiming at refining personalized treatment strategies in breast cancer care.

D. Feature Scaling

Furthermore, data has been divided into train and test groups. Then, using feature scaling, various units and magnitudes has been converted to at least one unit. Following this preprocessing, ML classifier has been used to find the simplest one. Numerous ML algorithms has been trained and and also tested in the dataset and discovered that the Bagging Classifier is the greatest fit, providing the highest accuracy rate.

E. Applying ML Classifier

Multiple algorithms have been applied in the project. For classification, KNN, SVC, Bagging Classifier and Adaboost Classifier has been used with the help of Random Forest as Base Regressor, Gradient Boosting, RandomForest and ANN. ANN does not give the best of result because data for training was not sufficient.

Whereas for Regression, SVR, Random Forest Regressor with Bagging and Adaboost on top of it, KNN, Decision Tree, ANN and Linear Regression has been used.

F. Model Training

A comprehensive dataset comprising data from 400 patients has been used in this study where each has been characterized by 117 distinct features. This rich dataset was instrumental in training the model, which was based on the Bagging classifier, a powerful machine learning algorithm known for its efficiency and accuracy in handling complex datasets. The primary focus of the model was on a binary target variable, represented by two classes: 0 and 1. This binary classification is aimed at differentiating specific patient outcomes or conditions.

IV. MODEL EVALUATION

A. Training, Testing, and Cross Validation

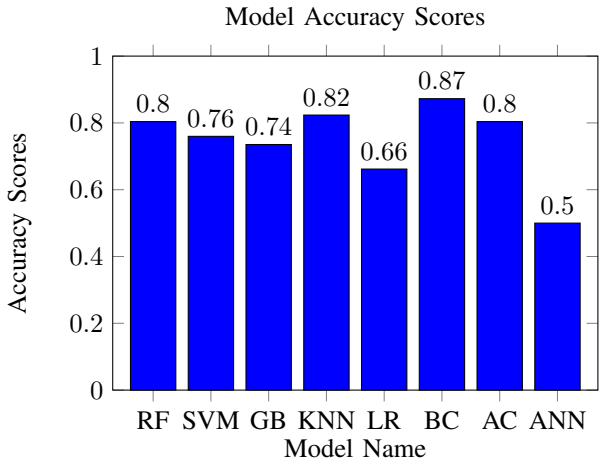
In this study, various machine learning models were rigorously trained, tested, and cross-validated to ensure effective breast cancer prognosis predictions. The dataset is divided with a ratio of 9:1 for the training and also testing purpose for the classification, whereas in the case of regression, the training and also testing are distributed in the ratio of 8:2, allowing a comprehensive learning phase while ensuring an unbiased evaluation. Models like Random Forest, SVM, and KNN were fine-tuned on the training data. The testing phase then assessed their predictive accuracy. Additionally, k-fold cross-validation was implemented, where the dataset was divided into multiple subsets, each serving as both a training and a testing set in separate iterations. For the classification, the bagging classifier gave the best performance, achieving a score of 87.25. In the case of regression, Gradient Boosting model demonstrates the best performance, achieving an MAE of 21.47.

B. Model Performance

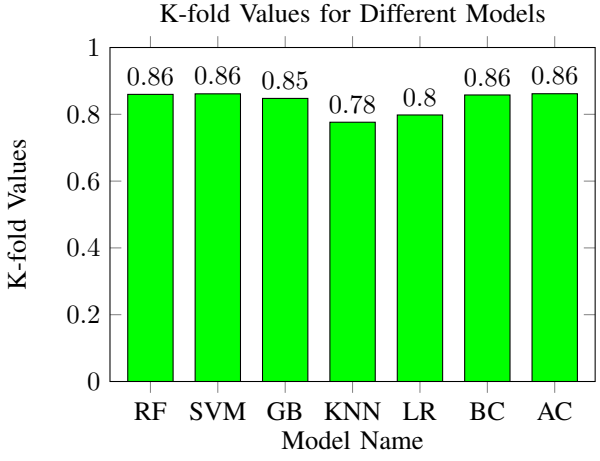
Based on the code analysis, model efficacy was quantitatively assessed across classification and regression tasks. Classification models, including Random Forest, SVM, and Gradient Boosting, demonstrated robust predictive performance, evidenced by high accuracy scores. In regression, models were evaluated using Mean Absolute Error (MAE), reflecting their reliability in forecasting continuous outcomes. This multifaceted evaluation approach provided a detailed insight into the models' predictive strengths and potential areas for optimization in breast cancer prognosis.

V. RESULTS AND ANALYSIS

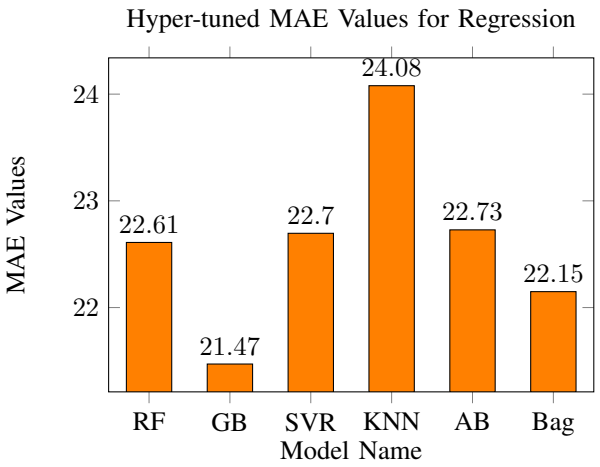
The balanced accuracy scores reveal different levels of performance among classification models. The Bagging Classifier achieves the highest score of 0.8725, demonstrating superior sensitivity and specificity balance. Random Forest and Adaboost Classifier are close behind with 0.8039 scores, indicating effective classification. KNN has a notable performance of 0.8235, indicating its ability to capture data patterns.



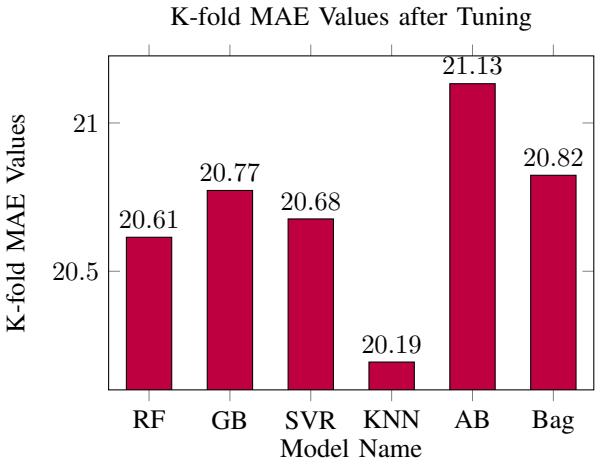
While performing K-Fold cross validation we saw that all the methods perform significantly well on the training data. The results of K-fold cross-validation show the model's performance and stability across multiple folds.



SVM and Adaboost Classifier perform consistently, with balanced accuracies of 0.8613 and 0.8616, respectively, demonstrating dependability. While Random Forest and Bagging Classifier perform well as well, KNN has slightly higher variability. The analysis aids in the selection of models with high accuracy and robustness for use in real-world scenarios.



The mean absolute error (MAE) values of models after hyperparameter tuning reflect improved predictive accuracy. Gradient Boosting has the lowest MAE at 21.4709, demonstrating improved precision. The Random Forest and Bagging models also perform well, with MAEs of 22.6101 and 22.1492, respectively. The tuning process improves the predictive capabilities of the models, resulting in more accurate and reliable predictions.



Model's mean absolute error (MAE) values after tuning demonstrate their improved accuracy. KNN has the lowest MAE of 20.1939, indicating better accuracy precision. The Random Forest and Bagging models also perform well, with MAEs of 20.6146 and 20.8236, respectively. AdaBoost has a slightly higher MAE of 21.1323, while Gradient Boosting and SVR perform similarly with MAEs of 20.7724 and 20.6764, respectively. The tuning process improves the predictive capabilities of the models, allowing for more accurate and reliable predictions.

VI. PROS AND CONS OF THE CHOSEN METHOD

With a mean value of 0.8579 and the best balanced accuracy score of 0.8725, the Bagging Classifier is the best option among the tables shown. It also performs consistently in K-fold cross-validation. Several noteworthy benefits of this model include its strong predictive power and less overfitting achieved by using ensemble methods. The training of numerous models can lead to greater computing effort, and the models may be less interpretable than simpler models like logistic regression. In spite of these drawbacks, the Bagging Classifier offers an appealing approach that successfully balances generalisation and accuracy—a critical balance for classification tasks that is related to the prognosis of this cancer.

Among the models assessed in the table, "Gradient Boosting" comes out as the most favourable classifier, with the lowest Mean Absolute Error (MAE) value of 21.4709. The accuracy of this model in predicting PCR and Relapse-Free Survival RFS is aided by an ensemble learning strategy that systematically corrects flaws of prior models. Despite its computational complexity and susceptibility to noisy data, Gradient Boosting excels at capturing subtle, nonlinear relationships within the dataset. However, users should be aware of the trade-off between computing needs and model interpretability, since the intricacy of Gradient Boosting may make it difficult to comprehend predictive aspects in medical diagnostics applications.

VII. CONCLUSION

This study effectively demonstrated the strong potential of ML algorithms to improve the prognostic precision of breast cancer by employing an extensive dataset from the I-SPY 2 TRIAL that included a range of MRI-based and clinical variables. With a balanced accuracy score of 87.25%, the Bagging Classifier was the best-performing model for classification; on the other hand, the Gradient Boosting method performed exceptionally well in regression tasks, with a Mean Absolute Error of 21.45. These findings contribute greatly to prediction reliability by demonstrating not only the effectiveness of the selected models but also the significance of tackling issues like feature selection and class imbalance using methods like SMOTE and Recursive Feature Elimination. The study highlights the potential of machine learning (ML) in personalised treatment planning and medical diagnostics, providing insightful information for better clinical judgement in the management of breast cancer. Moreover, it creates opportunities for further study, especially in the investigation of deep learning uses in cancer. In summary, the application of cutting-edge ML algorithms to the prediction of PCR and RFS is a promising weapon in the prevention and cure of breast cancer. It encourages the use of more individualised and data-driven methods in healthcare to improve patient outcomes and the standard of treatment as a whole.

REFERENCES

[1] P. Kathale and S. Thorat, "Breast Cancer Detection and Classification," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.367.

[2] C. Dubey, N. Shukla, D. Kumar, A. K. Singh and V. K. Dwivedi, "Breast Cancer Modeling and Prediction Combining Machine Learning and Artificial Neural Network Approaches," 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 119-124, doi: 10.1109/ICCCIS56430.2022.10037709.

[3] E. A. Bayrak, P. Kırcı and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 2019, pp. 1-3, doi: 10.1109/EBBT.2019.8741990.

[4] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, India, 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.

[5] 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.

[6] Anshuman and U. Kumar, "Machine Learning model for detection of Breast Cancer," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-4, doi: 10.1109/ISCON52037.2021.9702416.

[7] M. Akhil and P. V. S. Kumar, "Breast Cancer Prognosis using Machine Learning Applications," 2022 4th International Conference on Advances in Computing, Communication, Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 488-493, doi: 10.1109/ICAC3N56670.2022.10074517.

[8] V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2020, pp. 1-5, doi: 10.1109/C2I451079.2020.9368911.

[9] Y. Tewari, E. Ujjwal and L. Kumar, "Breast Cancer Classification Using Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 01-04, doi: 10.1109/ICACITE53722.2022.9823932.

[10] H. Sharma, P. Singh and A. Bhardwaj, "Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques," 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2022, pp. 1-6, doi: 10.1109/ESCI53509.2022.9758188.

VIII. CONTRIBUTION TABLE

Task and Weighting	*Data Pre-processing (10%)*	*Feature Selection (25%)*	*ML Method Development (25%)*	*Method Evaluation (10%)*	*Report Writing (30%)*	*Total*
Aditya Purswani 20596344	*10*	*20*	*20*	*30*	*20*	*100*
Anurag Phukan 20520078	*30*	*20*	*20*	*10*	*20*	*100*
Anasuya Dutta 20594248	*10*	*20*	*20*	*30*	*20*	*100*
Dhruv Bhattarcharjee 20592268	*30*	*20*	*20*	*10*	*20*	*100*
Shashwat Sinha 20520082	*20*	*20*	*20*	*20*	*20*	*100*
Total	*100*	*100*	*100*	*100*	*100*	

Fig. 4: Contribution Table of Group Members