IMPROVED BREAST CANCER DETECTION USING MACHINE LEARNING

AKASH M B[1], NITHEESH L[2], SHASHANK H U[3], SPOORTHY P[4], DR. ARUN NAGARLE[5]

Department of Information Science and Engineering Vidya Vikas Institute of Engineering and Technology [VVIET]

Mysuru, Karnataka, India

---***---

Abstract - This project investigated the application of machine learning to breast cancer detection. Logistic regression, K-nearest neighbors, Random Forest, and decision tree classifiers were implemented on a dataset of breast biopsy samples. Logistic regression and random forest achieved the highest accuracy (98.25% and 96.49%, respectively) in classifying malignant and benign cases. This project highlights the potential of machine learning in breast cancer detection and recognizes the need for further exploration of feature engineering and model optimization techniques. Future efforts will focus on improving the generalizability, interpretability, and verifiability of the model in clinical practice.

***Keywords**: Breast Cancer, Machine Learning, Classification, Early Detection, Logistic Regression, Random Forest.*

1. INTRODUCTION

Breast cancer is a prevalent global health concern, affecting millions annually. It manifests in various forms, ranging from non-invasive to invasive, with the potential to spread if left unchecked. Beyond its physical impact, breast cancer can significantly affect patients and their families emotionally and psychologically. Early detection is paramount in improving breast cancer outcomes, with significantly higher success rates for treatment and survival when identified early. Established screening methods like mammography, clinical breast exams, and self-examination empower individuals to detect abnormalities at an early stage, enabling timely medical intervention. However, despite advancements in screening and treatment, challenges persist in ensuring widespread access to early detection and quality healthcare for all at risk or affected by breast cancer. To address these concerns, a multi-faceted approach is necessary, encompassing awareness campaigns, education, improved screening technologies, and equitable access to healthcare resources. In this context, machine learning (ML) models for breast cancer detection emerge as a promising avenue for enhancing early detection efforts. By harnessing the power of computational algorithms and vast datasets, these models can assist healthcare professionals in accurately identifying and classifying breast abnormalities. This, in turn, facilitates timely diagnosis and treatment planning, potentially improving patient outcomes.

The paper investigates how machine learning can be used to identify breast cancer. Our objective is to develop robust and accurate models that can aid healthcare practitioners in their clinical decision-making processes. Through this endeavor, we aim to contribute to the ongoing fight against breast cancer and improve outcomes for individuals affected by this disease.

2. LITERATURE REVIEW

Worldwide, breast cancer is the primary cause of death for women. Improved survival rates and effective treatment depend on early detection. Machine learning (ML) is quickly becoming a potent technology for highly accurate and efficient breast cancer prediction. Examining the state of research in this field at the moment, this review highlights important discoveries, useful techniques, and exciting new avenues for further investigation. This review used terms like "breast cancer prediction" and "machine learning" to conduct a thorough search of scholarly databases like PubMed, Scopus, ResearchGate, and Google Scholar. A review was conducted on a subset of selected articles based on their methodology, findings, relevance, and field contribution. A short life expectancy, severe disease, and frailty are possible outcomes of breast cancer. Furthermore, it could also be lethal. Therefore, it may be said that breast cancer is now a serious worry. This study focuses on machine learning methods that aid in identifying and perceiving multiple breasts. Several machine learning approaches are covered here, such as feature selection, support vector machines, hidden Markov models, data mining techniques, genetic algorithms, prediction systems, and computational intelligent classifiers. Benefit Machine learning techniques perform well when handling multidimensional and multivariate data in dynamic or uncertain environments. One drawback of machine learning is the requirement for training big, comprehensive, unbiased data sets of excellent quality.

3. PROBLEM STATEMENT

Breast cancer poses a significant global health burden, ranking as the most prevalent cancer among women worldwide (GLOBOCAN, 2020). In 2020 alone, it resulted in over 2.3 million diagnoses and 685,000 deaths, highlighting its devastating impact (GLOBOCAN, 2020). While the exact causes often remain unclear, DNA damage within breast cells plays a key role. Established risk factors include age, obesity, alcohol consumption, family history, radiation exposure, reproductive history, smoking, and postmenopausal hormone therapy (American Cancer Society, [Year]). However, a concerning number of cases, exceeding 50%, occur in women with no identifiable risk factors beyond being over 40 years old (American Cancer Society, [Year]). Genetic mutations, particularly in BRCA1, BRCA2, and PALB2 genes, further elevate risk (National Cancer Institute, [Year]).

The burden of breast cancer is particularly acute in India, where it accounted for 13.5% of all cancers and 10.6% of cancer deaths in 2020 (GLOBOCAN, 2020). In 2018, India witnessed a staggering 162,468 new cases and 87,090 deaths (GLOBOCAN, 2020). Over half of Indian women are diagnosed at advanced stages (stage 3 or 4), significantly hindering treatment options and reducing survival rates by 60%

(reference on advanced stage diagnosis in India). Delayed diagnoses and complex treatment regimens contribute to these diminished survival rates. Early detection offers the best chance for successful outcomes, yet access to timely screening remains a challenge, particularly in resource-constrained settings (reference on screening disparities in India). Addressing the global challenge of breast cancer necessitates heightened awareness, improved early detection strategies, and ensuring equitable access to treatment. Through these combined efforts, we can strive to improve survival rates and overall outcomes for women battling breast cancer worldwide.

4. METHODOLOGY

This study investigated the efficacy of various machine-learning algorithms for breast cancer detection using a publicly available dataset of breast tissue biopsies. The dataset contained features extracted from digitized images, including measurements of cell nuclei size, shape, and texture.

Data Preprocessing

- Data exploration revealed dataset structure and basic statistics.
- Irrelevant columns ("id", "Unnamed: 32") were removed.
- The target variable ("Diagnostic") indicating malignancy ("M") or benignity ("B") was converted to binary (1 for malignant, 0 for benign) for compatibility with machine learning models.
- A summary table visualized the distribution of malignant and benign cases, highlighting potential class imbalances.
- A heatmap of the correlation matrix explored relationships between features, aiding in feature selection.

Data Splitting and Standardization

- The dataset was divided into training (80%) and testing (20%) sets.
- Feature scaling (Standardization) ensured all features had a mean of 0 and standard deviation of 1, promoting equal contribution to model training and improved convergence, especially for gradient-based algorithms.

Model Building and Evaluation

- Four algorithms were employed: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Decision Tree.
- Each model was trained on the standardized training data and evaluated on the testing set.

Model Descriptions

- **Logistic Regression:** A linear model adept at binary classification tasks. It estimates the probability of a data point belonging to a specific class (malignant or benign) based on its features.
- **KNN:** This algorithm classifies data points by considering the majority class of their K-nearest neighbors within the feature space. The optimal value of K, the number of neighbors considered, was determined through experimentation**.**
- **Random Forest:** An ensemble learning method that leverages the power of multiple decision trees. During training, a random forest constructs numerous decision trees, and the final classification is based on the most frequent class predicted by these individual trees.
- **Decision Tree:** A non-parametric method that builds tree-like decision models by recursively splitting data based on feature values. It is interpretable and easy to visualize.

Model Performance Evaluation

- **Confusion Matrix:** Provided insights into true positives (correctly classified malignant cases), true negatives (correctly classified benign cases), false positives (incorrectly classified malignant - benign cases), and false negatives (incorrectly classified benign - malignant cases). A heatmap was generated for better visualization.
- **AUC - ROC scores** were utilized to visually assess model performance. ROC curves plot the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. This visualization allows for a clear comparison of how well each model distinguishes between positive (malignant) and negative (benign) cases. The AUC score summarizes the overall performance of a model by quantifying the area under the ROC curve. A higher AUC score indicates a greater ability of the model to differentiate between the two This comprehensive evaluation process enabled the identification of the most effective breast cancer detection algorithm for the specific dataset employed in this study. It is important to note that algorithm performance can vary depending on the characteristics of the data.

5. RESULT

This study used a specific breast cancer detection dataset to assess the performance of different classification algorithms for breast cancer diagnosis. Four machine learning models were examined in terms of performance: Decision Tree, Random Forest, K-Nearest Neighbors (KNN) Classifier, and Logistic Regression. For every model, a thorough assessment methodology that included accuracy, classification report, and confusion matrix analysis was put into place. These measures offered insightful information on how well the models classified cases of benign and aggressive breast cancer.

Logistic Regression:

- Confusion Matrix: Successfully classified all 71 benign cases and accurately classified 41 out of 43 malignant cases.
- Accuracy: Achieved an impressive accuracy of 98.25%, indicating high correctness in classifying both malignant and benign cases.
- Classification Report: Demonstrated high precision, recall, and F1-score for both malignant and benign classes, indicating balanced performance across evaluation metrics.

KNN Classifier:

- Confusion Matrix: Accurately classified 70 out of 71 benign cases and 36 out of 43 malignant cases, with some misclassifications.
- Accuracy: Achieved a slightly lower accuracy of 92.98% compared to Logistic Regression.
- Classification Report: While performance metrics for benign cases were high, there was a slight decrease in performance for malignant cases, resulting in lower overall accuracy compared to Logistic Regression.

Random Forest Classifier:

- Confusion Matrix: Successfully classified 70 out of 71 benign cases and 40 out of 43 malignant cases, with very few misclassifications.
- Accuracy: The model achieved a commendable accuracy of 96.49%, demonstrating its effectiveness in breast cancer classification.
- Classification Report: The model exhibited high precision, recall, and F1-score for both malignant and benign classifications. This balanced performance across classes indicates its robustness in identifying both positive and negative cases of breast cancer.

Decision Tree Classifier:

- Confusion Matrix: Correctly classified 68 out of 71 benign cases and 37 out of 43 malignant cases, with some misclassifications.
- Accuracy: The model reached 92.11% accuracy, better than KNN but lower than Random Forest.
- Classification Report: Showed relatively high precision, recall, and F1-score for both malignant and benign classes, indicating satisfactory performance.

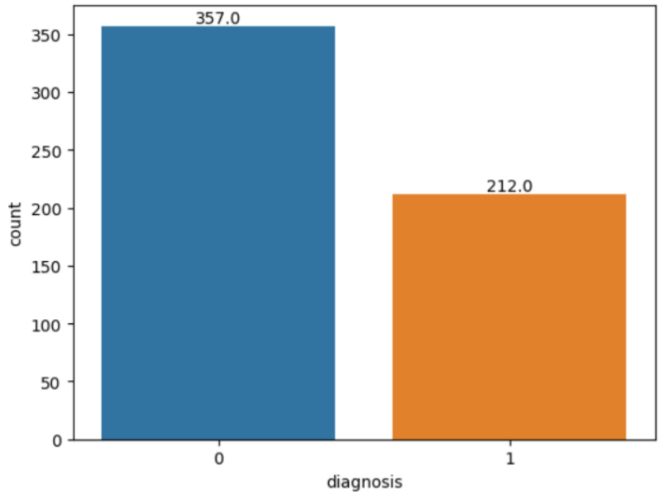

Fig: Count Plot

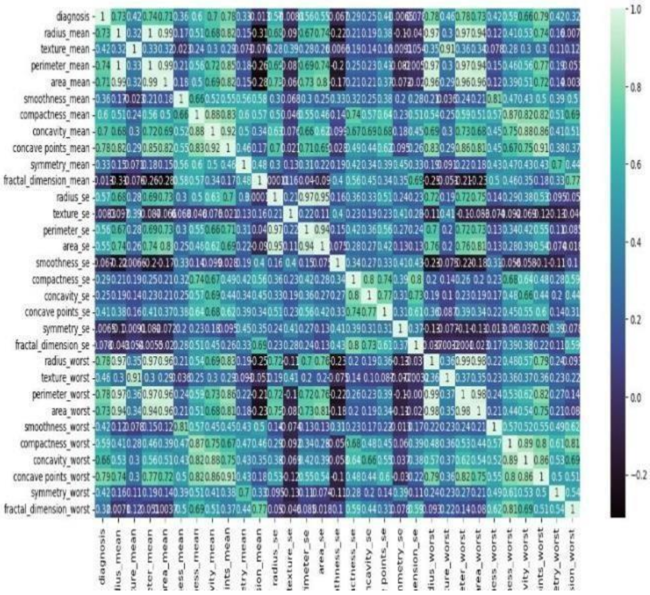

Fig: Heat Map

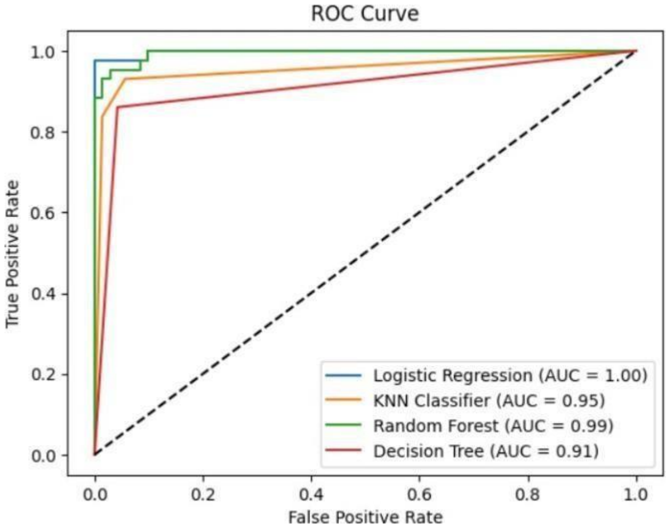

Fig: ROC -AUC Curve

5. CONCLUSIONS

In summary, the research aimed to diagnose breast cancer by applying machine learning methods to create prediction models that might identify cases of the disease as benign or malignant by analyzing imaging and clinical data. The study yielded important insights and results through extensive data preprocessing, exploratory data analysis, model training, and evaluation. An overview of the project is as follows: The project's goal was to use machine learning algorithms to help with breast cancer early identification and detection, a major global healthcare issue. The project aimed to develop an accurate and dependable predictive mode by analyzing ten real-valued features (such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) computed for each cell nucleus extracted from breast biopsy samples.

6. REFERENCES

1. Zhong, X. et al. (2023) 'Using machine learning to predict diagnosis and survival outcomes for bone metastasis in breast cancer.'
2. Maniruzzaman, Md. et al. (2018) 'Utilizing machine learning for accurate diabetes risk stratification: The impact of missing values and outliers.'
3. Rasool, A. et al. (2022) 'Enhanced predictive models for breast cancer diagnosis using machine learning,' International Journal of Environmental Research and Public Health.
4. Aamir, S. et al. (2022) 'Leveraging supervised machine learning techniques to predict breast cancer,' Computational and Mathematical Methods in Medicine.
5. Gigi F. Stark, Gregory R. Hart, Bradley J. Nartowt, Jun (2022) 'Predicting breast cancer risk using personal health data and machine learning models.'