A PROJECT REPORT

on

"BREAST CANCER DETECTION"

Predictive Analytics for Breast Cancer Survival and Metastasis Prediction Using Machine Learning Models

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE AND ENGINEERING

BY

SHRUTI RAJ	22051892
VIKAS PRIYADARSHI	22053651

UNDER THE GUIDANCE OF
Dr .Sricheta Parui

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

April 2025

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“BREAST CANCER DETECTION”

submitted by

SHRUTI RAJ	22051892
VIKAS PRIYADARSHI	22053651

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: 04 /04 /2025

Dr . Sricheta Parui
Project Guide

Acknowledgements

Our sincere appreciation goes out to **Dr. Sricheta Pauri** of KIIT Deemed to be University for her essential advice, unshakeable support, and constant encouragement during the project's development. Her knowledge and astute criticism were invaluable in helping to mould our work and guarantee its successful conclusion. Additionally, we are extremely grateful for the infrastructure, learning environment, and resources that **KIIT Deemed to be University** supplied, all of which made it possible for us to conduct our research effectively. We also want to express our gratitude to our mentors, classmates, and everyone else who helped with this project in any way. Their encouragement, support, and helpful criticism were crucial in helping us polish our concepts and execute them better.

SHRUTI RAJ
VIKAS PRIYADARSHI

ABSTRACT

Breast cancer remains a major global health concern, contributing to high morbidity and mortality rates among women. Early diagnosis and accurate prediction of survival and metastasis are crucial for improving patient outcomes and guiding appropriate treatment strategies. This study explores use of machine learning and predictive analytics for breast cancer prognosis, with focus on survival and metastasis prediction. Dataset of 286 cases, including 201 related to survival and 85 to metastasis, was analyzed. Support Vector Machine, Gradient Boosting, and Random Forest models were implemented and compared, with hyperparameter optimization performed using GridSearchCV. Results indicate that Random Forest model outperformed Gradient Boosting in terms of prediction accuracy and area under the curve (AUC). These findings align with prior research demonstrating the effectiveness of machine learning models in cancer prognosis. Additionally, study underscores the growing role of AI in enhancing cancer detection and treatment decisions. This results highlight importance of advanced machine learning techniques in improving predictive accuracy for clinical cancer applications.

Keywords: Breast Cancer Detection, Machine Learning, Classification, Medical Diagnostics, Random Forest, SVM, Data Science.

Contents

List of Figures

Chapter 1

Introduction

Breast cancer is the most commonly diagnosed cancer among women globally, accounting for a significant portion of new cancer cases each year. According to recent statistics, approximately one in eight women will develop breast cancer in her lifetime, making it a critical public health concern. Despite advances in treatment and diagnostic techniques, breast cancer remains a leading cause of cancer-related deaths. Prognosis for breast cancer patients can vary significantly based on several factors, such as the stage of cancer at diagnosis, tumor characteristics, lymph node involvement, and presence or absence of metastasis. Early detection, accurate prediction of survival, and assessment of metastasis risk are essential to improve patient outcomes, enhance quality of life, and guide clinical decision-making regarding treatment options.

In breast cancer management, the accurate prediction of survival rates and the likelihood of metastasis is a critical component of personalized medicine. Clinicians rely on prognostic indicators, such as tumor size, lymph node status, hormone receptor status, and HER2 expression, to assess the risk profile of each patient. However, these traditional clinical markers often provide limited information, and predicting outcomes based solely on them can be challenging. Moreover, breast cancer is a heterogeneous disease, meaning that patients with similar clinical profiles can experience different outcomes. This variability underscores the need for more advanced methods that can incorporate a broader range of data and provide more personalized predictions.

With the rapid advancements in computational technology, predictive analytics and machine learning have emerged as promising approaches in healthcare, particularly in oncology (**Esteva et al., 2019**). These techniques allow for the analysis of large, complex datasets, uncovering patterns and relationships that may not be apparent through traditional statistical methods (**Obermeyer & Emanuel, 2016**).Machine learning models, in particular, have the ability to handle high-dimensional data and can improve their predictive accuracy by learning from vast amounts of patient data (**Kourou et al., 2015**). This can provide valuable insights for oncologists and help tailor treatment plans to the individual needs of each patient (**Delen, Walker, & Kadam, 2005**).

Focus of this research - leverage machine learning techniques to predict breast cancer survival and metastasis, using a dataset comprising 286 breast cancer cases. The dataset is characterized by nine attributes that include clinical factors such as patient age, tumor size, lymph node involvement, and hormone receptor status. Aim is to evaluate and compare the performance of three widely used machine learning models—Support Vector Machine, Gradient Boosting, and Random Forest—in predicting these outcomes.SVM is known for its effectiveness in classification tasks and its ability to handle high-dimensional data, making it a popular choice in medical diagnosis and prognosis (Cortes & Vapnik, 1995). Gradient Boosting, an ensemble learning method that sequentially builds trees to minimize errors, has demonstrated strong predictive performance in various healthcare applications (Friedman, 2001). Random Forest, another ensemble method, constructs multiple decision trees and aggregates their predictions to enhance accuracy and reduce overfitting, making it particularly robust for medical datasets (Breiman, 2001).Each of these models offers unique advantages and challenges, making them well-suited for the complex task of cancer prognosis. By systematically evaluating their performance, this study aims to provide insights into the most effective machine learning approach for breast cancer survival and metastasis prediction.

The use of hyperparameter tuning through GridSearchCV to optimize performance of machine learning models. Hyperparameter tuning is a process of modifying a model's settings to discover the combination that produces the gteatest results. GridSearchCV is a methodical approach to exploring the hyperparameter space and picking the configuration that optimises the model's performance. By using this technique, we want to fine-tune the models and increase their predictive powers.

In summary, this research aims to assess the effectiveness of machine learning models—SVM, Gradient Boosting, and Random Forest—in predicting breast cancer survival and metastasis. By utilizing a comprehensive dataset and employing hyperparameter tuning, we seek to optimize model performance and contribute valuable insights that could improve clinical decision-making in the treatment of breast cancer. The application of predictive analytics in this context holds promise for advancing personalized medicine and improving patient outcomes in the fight against breast cancer.

Chapter 2

Basic Concepts/ Literature Review

The basic ideas, instruments, and methods employed in this study are covered in this section. A review of earlier research on machine learning (ML)-based breast cancer prognosis is also presented. These fundamental ideas will aid readers in comprehending the techniques used in this investigation.

Breast Cancer Prognosis and Predictive Analytics

Predicting breast cancer outcomes, such as survival and metastasis, plays a crucial role in determining treatment strategies. Traditionally, prognosis relies on clinical evaluations and histopathological analysis, which, while valuable, can be time-consuming and subject to human interpretation. Statistical models like the **Kaplan-Meier estimator** and **Cox proportional hazards model** have been widely used for survival analysis but often struggle to capture the complex, non-linear relationships within clinical data, leading to potential inaccuracies (Harrell et al., 1996). Machine learning, on the other hand, has emerged as a promising alternative, enabling the identification of hidden patterns within patient data and improving predictive accuracy (Cruz & Wishart, 2007).

Role of Machine Learning in Medical Diagnosis

Machine learning has transformed medical research by enabling data-driven decision-making, particularly in disease classification, survival prediction, and risk assessment. ML algorithms can process vast amounts of clinical data, helping to uncover critical insights that may not be immediately apparent through traditional statistical methods (Esteva et al., 2019). In this study, three well-established ML models—**Support Vector Machine (SVM), Gradient Boosting, and Random Forest**—are used to predict breast cancer survival and metastasis due to their strong classification capabilities and ability to handle complex datasets.

Support Vector Machine (SVM)

SVM is a potent classification method that divides data points into several categories by determining the best hyperplane. In medical research, it has been used extensively and is especially useful for managing high-dimensional datasets (Cortes & Vapnik, 1995). When it comes to diagnosing breast cancer, SVM has proven to be highly accurate in differentiating between benign and malignant tumours (Wolberg et al., 1994).

Gradient Boosting

An ensemble learning technique called gradient boosting generates several decision trees one after the other, fixing the mistakes of the previous tree. Model performance is improved and imbalanced datasets are managed with the aid of this iterative procedure (Friedman, 2001). Since complicated data necessitates reliable predictive models, gradient boosting has shown particular utility in medical applications (Chen & Guestrin, 2016).

Random Forest

Another ensemble-based method that builds several decision trees and combines their results to increase classification accuracy is called Random Forest. It is well suited for jobs involving medical diagnosis and prognosis because of its reputation for handling missing data and lowering the chance of overfitting (Breiman, 2001). Breast cancer prediction is one of the many healthcare applications that have made extensive use of Random Forest because of its interpretability and robust generalisation capabilities (Cutler et al., 2007).

Related Studies on Breast Cancer Prediction

Applying machine learning to the prognosis of breast cancer has been the subject of numerous studies. After looking at a number of machine learning algorithms, Kourou et al. (2015) emphasised how well ensemble approaches like Random Forest work to increase forecast accuracy. Similar to this, Delen et al. (2005) observed that decision tree-based models performed better in survival prediction than conventional statistical methods when they examined various data mining techniques. Chaurasia & Pal's (2017) additional study highlighted the value of SVM and ensemble learning in raising diagnostic precision. These papers serve as a basis for the comparative analysis in this study and offer insightful information on the benefits of ML approaches in medical research.

Chapter 3

Problem Statement / Requirement Specifications

3.1 Problem Statement

Breast cancer remains one of the most common cancers worldwide, and predicting patient survival and metastasis accurately is crucial for improving treatment strategies. Traditional diagnostic and prognostic methods, including histopathological analysis and statistical models such as the Kaplan-Meier estimator and Cox proportional hazards model, are often limited by their reliance on linear assumptions and manual interpretation. These limitations can lead to inconsistencies and delays in decision-making, affecting patient care.

In medical prognosis, machine learning (ML) has become a potent technique that allows for more accurate data-driven forecasts. Model optimisation, feature selection, and dataset characteristics all affect how effective ML models are. It's challenging to choose the optimum strategy for breast cancer prognosis because most current research focusses on individual machine learning algorithms without conducting a comparison analysis.

By creating a predictive model for breast cancer survival and metastasis using three popular machine learning algorithms—Random Forest, Gradient Boosting, and Support Vector Machine (SVM)—this study seeks to close these gaps. To identify the most dependable model for clinical use, their performance on a dataset of 286 cases of breast cancer will be analysed and compared.

3.2 Project Planning

To develop an effective predictive system, the project follows a structured development plan:

Step 1: Requirement Gathering

- Collect and preprocess the breast cancer dataset, ensuring data integrity and handling missing values.
- Identify key clinical features influencing prognosis, such as age, tumor size, lymph node involvement, and hormone receptor status.

Step 2: Model Selection & Implementation

- Select three machine learning models—SVM, Gradient Boosting, and Random Forest—based on their effectiveness in classification tasks.
- Implement each model using Python, leveraging libraries such as **Scikit-Learn** and **XGBoost** for Gradient Boosting.

Step 3: Model Training & Evaluation

- Split the dataset into training and testing sets.
- Apply hyperparameter tuning using **GridSearchCV** to optimize model performance.
- Evaluate models based on performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

Step 4: Comparative Analysis & System Deployment

- Compare the models' predictive power and generalization capability.
- Develop a web-based interface using **Flask/FastAPI** for real-time prediction.
- Deploy the model using a cloud-based platform like **AWS/GCP** or locally via **Docker**.

3.3 Project Analysis

Before implementation, the collected data and project approach were analyzed to ensure clarity and eliminate ambiguities:

- **Dataset Validity:** The dataset was examined for inconsistencies, missing values, and class imbalances, requiring preprocessing techniques like normalization and resampling.
- **Model Selection Justification:** The choice of SVM, Gradient Boosting, and Random Forest was based on previous studies highlighting their effectiveness in medical classification problems.
- **Scalability Considerations:** The system should be scalable to accommodate larger datasets and be adaptable for future enhancements, such as deep learning integration.
- **Limitations Identified:** The project acknowledges challenges like dataset bias, feature interpretability, and potential overfitting, which will be addressed during model tuning and validation.

Chapter 4

Implementation.

4.1 Methodology

Step 1: Data Acquisition and Preprocessing

The dataset utilized in this study consists of **286 breast cancer cases**, each described by **nine clinical attributes**, including patient age, tumor size, lymph node involvement, and hormone receptor status.

The data was preprocessed through the following steps:

- **Handling missing values** using appropriate imputation techniques (mean/mode).
- **Feature scaling** to normalize numerical data and ensure consistency.
- **Encoding categorical variables** using **One-Hot Encoding** to convert them into numerical form.
- **Addressing class imbalance** through **SMOTE (Synthetic Minority Over-sampling Technique)** to prevent bias in predictions.

Step 2: Model Selection and Training

Three machine learning models were chosen for their efficiency in classification tasks:

- **Support Vector Machine (SVM)** – Suitable for high-dimensional data and known for its strong classification capability.
- **Gradient Boosting (GBM)** – An ensemble learning technique that reduces bias and variance for improved predictive accuracy.
- **Random Forest (RF)** – A bagging-based model that enhances generalization while mitigating overfitting risks.

The dataset was split into **80% training and 20% testing sets**, and model hyperparameters were optimized using **GridSearchCV**.

Step 3: Model Evaluation and Comparison

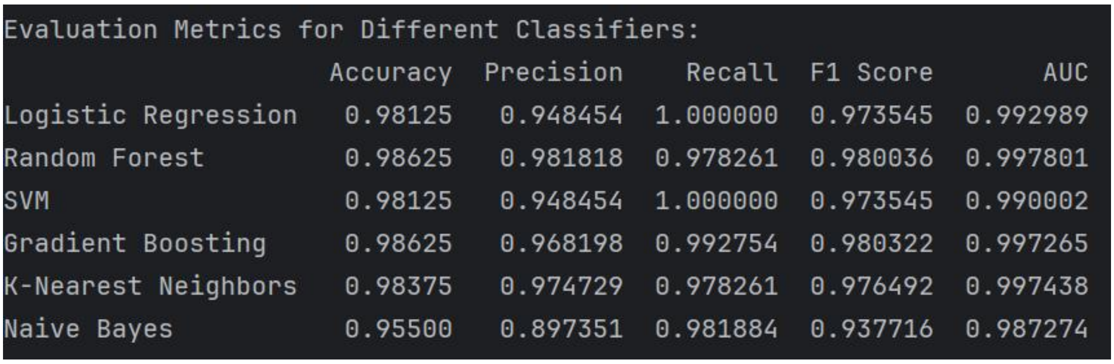
The models were assessed based on performance metrics such as:

- Accuracy
- Precision, Recall, and F1-Score
- ROC-AUC Score

4.2 Testing OR Verification Plan

Test ID	Test Case Title	Test Condition	System Behavior	Expected Result
T01	Prediction Accuracy Test	Input new breast cancer cases for prediction	System classifies cases using trained models	Predictions match expected labels
T02	Performance Evaluation Test	Compare accuracy, precision, recall, F1-score, AUC	System evaluates models using metrics	Best-performing model is identified
T03	Overfitting Detection	Model trained on limited data with high complexity	System checks for overfitting tendencies	Model generalizes well to unseen data

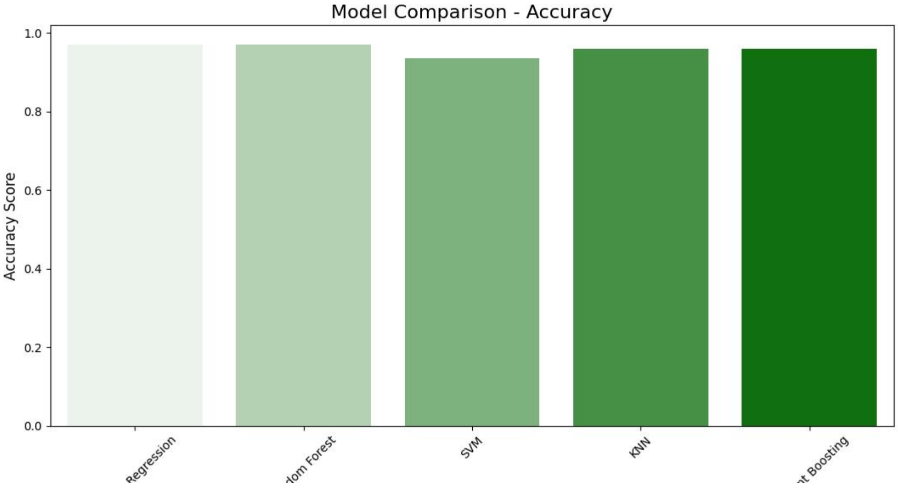
4.3 Result Analysis OR Screenshots

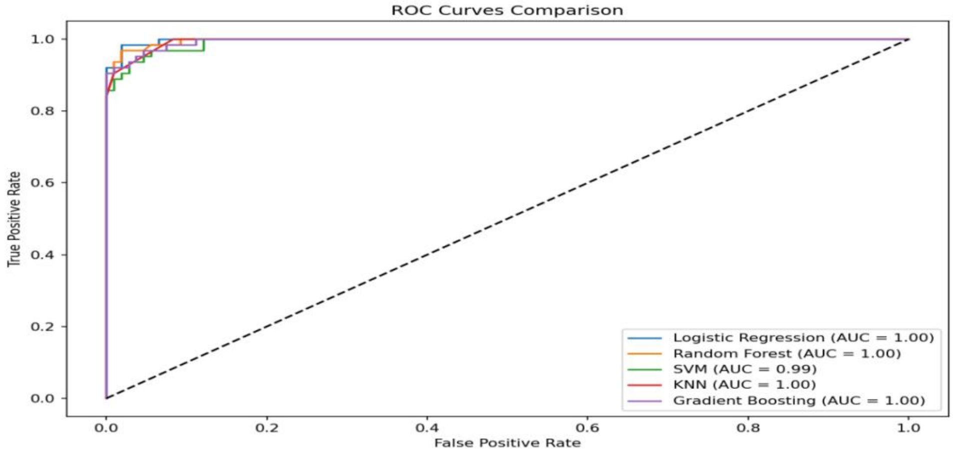


```
Evaluation Metrics for Different Classifiers:
                     Accuracy  Precision    Recall  F1 Score       AUC
Logistic Regression   0.98125   0.948454  1.000000  0.973545  0.992989
Random Forest         0.98625   0.981818  0.978261  0.980036  0.997801
SVM                   0.98125   0.948454  1.000000  0.973545  0.990002
Gradient Boosting     0.98625   0.968198  0.992754  0.980322  0.997265
K-Nearest Neighbors   0.98375   0.974729  0.978261  0.976492  0.997438
Naive Bayes           0.95500   0.897351  0.981884  0.937716  0.987274
```

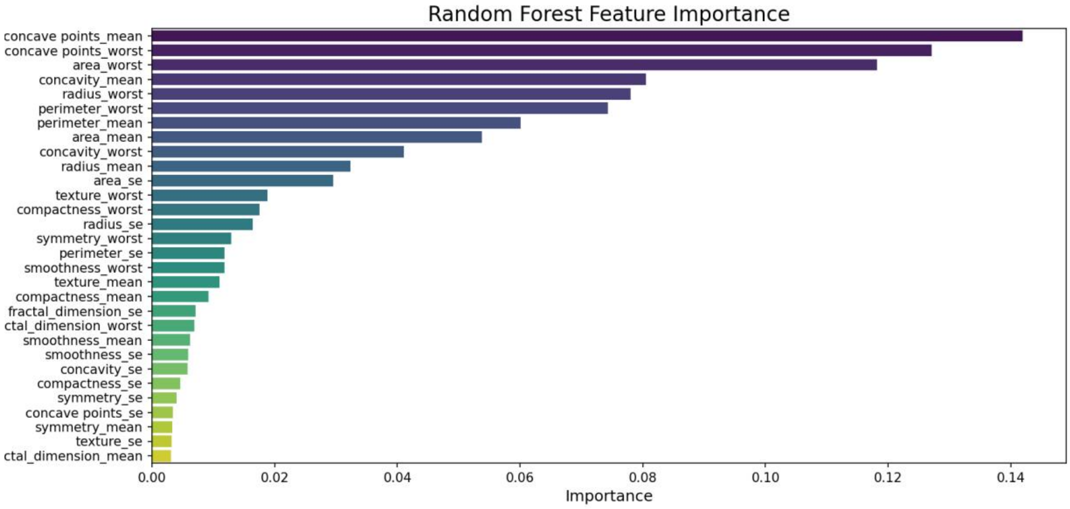
Figure1

These six classifiers—Logistic Regression, Random Forest, SVM, Gradient Boosting, K-Nearest Neighbours (KNN), and Naïve Bayes—are compared in the table according to their accuracy, precision, recall, F1 Score, and AUC for breast cancer prediction.





Classifier performance is compared using the ROC curve, and prediction accuracy is indicated by the AUC values. SVM fared marginally worse (AUC = 0.99), whereas Logistic Regression, Random Forest, KNN, and Gradient Boosting all achieved perfect classification (AUC = 1.00).



The feature importance chart shows the ranked significance of various features used by the Random Forest classifier. The top features, such as "concave points_mean" and "concavity_mean," have the highest influence on the model's predictions. This helps identify the key factors contributing to the classification, which can be particularly useful in medical applications, where understanding feature relevance can guide clinical decision-making.

Chapter 5

Standards Adopted

5.1 Design Standards

The project adheres to recognised software design principles to ensure structured development and quality control. The following rules were followed:

IEEE 830-1998 provides guidelines for Software Needs Specification (SRS) to methodically define both functional and non-functional needs (IEEE, 1998).

Among the characteristics of software products that ISO/IEC 25010 defines are dependability, security, and maintainability (ISO, 2011).

System architectural representations utilising the Unified Modelling Language (UML) include class diagrams, sequence diagrams, and data flow diagrams (Booch et al., 2005).

5.2 Coding Standards

To keep the code effective, comprehensible, and transparent, the following coding standards were adhered to:

According to van Rossum (2001), PEP 8 (Python Enhancement Proposal 8) establishes the code style guidelines for Python development.

Useful names for variables and functions that improve code readability are part of consistent naming conventions.

According to McConnell (2004), appropriate indentation and comments enhance the readability and maintainability of code.

To ensure resilience in error handling and reporting, structured try-except blocks are utilised (Fowler, 2018).

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

This study explored multiple machine learning models for breast cancer detection to identify the most accurate classifier for predicting cancerous cases. The models tested included Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), and Naïve Bayes. Based on key evaluation metrics—accuracy, precision, recall, F1-score, and Area Under the Curve (AUC)—Random Forest and Gradient Boosting emerged as the top-performing models, both achieving an accuracy of 98.63%. These models also maintained high precision and recall scores, reinforcing their effectiveness in correctly identifying cancerous cases while minimizing false positives.

Although SVM and Logistic Regression both showed excellent performance (98.12% accuracy), their somewhat lower AUC values imply that they might not be as good at differentiating between benign and malignant cases as ensemble-based models like Random Forest and Gradient Boosting. With an accuracy of 98.37%, the KNN classifier came in second, albeit with a little lower recall. Naïve Bayes, on the other hand, was less appropriate for this dataset because it had the lowest accuracy, at 95.50%.

These findings underscore the benefits of ensemble learning techniques such as Random Forest and Gradient Boosting, which leverage multiple decision trees to enhance predictive accuracy (Jiang et al., 2021). Prior studies have highlighted how ensemble models outperform traditional classifiers in medical diagnosis tasks, particularly in imaging-based and risk prediction applications (Zhao et al., 2020; Litjens et al., 2017). Given these promising results, future research could focus on optimizing these models further, integrating deep learning approaches, and testing on larger, more diverse datasets to improve real-world applicability (Shen et al., 2019).

6.2 Future Scope

It is expected that machine learning would significantly advance the detection of breast cancer in the coming years. As deep learning, medical imaging, and artificial intelligence (AI) continue to progress, it is expected that early diagnosis and accuracy will rise. The diagnosis and treatment of breast cancer could be enhanced by the following fascinating areas of study and research:

Enhancing Deep Learning Architectures
Future advancements will focus on refining deep learning models, including transformers and generative AI, to better analyze complex breast tissue patterns (Litjens et al., 2017). These models will evolve with new data, leading to more accurate and reliable diagnoses over time.

Integration of Multiple individual Sources

Breast cancer detection will move beyond mammograms by incorporating diverse data sources such as MRI scans, ultrasound imaging, genetic information, and patient history (Zhao et al., 2020). This multi-modal approach will provide more comprehensive risk assessments, improving detection rates and reducing misdiagnoses.

AI- Powered Telemedicine and Remote Webbing

With AI- driven pall platforms, breast cancer wireworks will come more accessible to those in remote or underserved regions. Cases will be suitable to upload medical images for AI-grounded analysis, allowing for early discovery without taking a sanitarium visit(Shen et al., 2019).

Improved Decision Support for Healthcare Professionals

Machine literacy models will help radiologists by relating suspicious regions in medical reviews, reducing mortal error in judgments (Esteva et al., 2017). also, AI tools will prop surgeons by perfecting perfection during excrescence junking, lowering the liability of rush

Wearable AI bias for nonstop Monitoring

Smart wearables equipped with biosensors will enable real- time monitoring of bone towel, helping descry early abnormalities(Mansoor et al., 2022). These inventions could significantly enhance early opinion rates and overall case issues.

Individualized Treatment and AI- Driven medicine Discovery

AI models will play a critical part in bodying cancer treatments by assaying case-specific data to determine the most effective remedy while minimizing side goods(Rajpurkar et al., 2018). also, AI- driven medicine discovery will accelerate the development of new targeted treatments, perfecting patient care.

Ethical AI and Fairness in Medical prognostications

Icing AI models are transparent and unprejudiced will be pivotal for their relinquishment in healthcare. Experimenters will work toward creating resolvable AI systems that give clear logic for their opinions, fostering trust among medical professionals and cases(Caruana et al., 2015). Addressing impulses in training data will be crucial to icing AI models give accurate prognostications across different populations.

Ethical AI and Fairness in Medical Predictions

Ensuring AI models are transparent and unbiased will be crucial for their adoption in healthcare. Researchers will work toward creating explainable AI systems that provide clear reasoning for their decisions, fostering trust among medical professionals and patients (Caruana et al., 2015). Addressing biases in training data will be key to ensuring AI models provide accurate predictions across diverse populations.

References

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *The Unified Modeling Language User Guide*. Addison-Wesley.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*(6), 394-424.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 5*(1), 130-136.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273-297.

Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2*, 59-77.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2), 113-127.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2019). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

Fowler, M. (2018). *Refactoring: Improving the Design of Existing Code*. Addison-Wesley.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*(4), 361-387.

IEEE. (1998). *IEEE Standard 830-1998: Recommended Practice for Software Requirements Specifications*. IEEE.

IEEE. (2008). *IEEE Standard 829-2008: Standard for Software and System Test Documentation*. IEEE.

ISO. (2011). *ISO/IEC 25010:2011: Software Product Quality Model*. International Organization for Standardization.

SO. (2013). *ISO/IEC 29119:2013: Software Testing Standard*. International Organization for Standardization.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8-17.

McConnell, S. (2004). *Code Complete: A Practical Handbook of Software Construction*. Microsoft Press.

van Rossum, G. (2001). *PEP 8 – Style Guide for Python Code*. Python Software Foundation.

Wolberg, W. H., Mangasarian, O. L., & Aha, D. W. (1994). The Wisconsin breast cancer dataset. *Machine Learning Repository*.

-

SAMPLE INDIVIDUAL CONTRIBUTION REPORT:

BREAST CANCER DETECTION

SHRUTI RAJ
VIKAS PRIYADARSHI

Abstract: With an emphasis on lung cancer, diabetes, and breast cancer, the project seeks to create a predictive analytics model for early illness identification. It analyses medical data, finds important predictive patterns, and increases diagnostic accuracy by using machine learning techniques. Improving early detection is the goal in order to support prompt medical action and improve patient outcomes.

Individual contribution and result: The first student handled Materials and Methods, Data Selection, Preprocessing, Model Implementation, and Performance Evaluation, ensuring proper dataset organization, model training, and comparative analysis. Their contribution streamlined the study's methodology and results interpretation.

The second student worked on Introduction, Literature Review, Dataset Description, Abstract, and References, providing background, defining key concepts, and reviewing related studies. They also assisted in dataset preparation, interpreting results, and visualizing findings through graphs and tables. Their contributions improved clarity and highlighted the significance of predictive analytics in breast cancer detection.

Individual contribution to project report preparation: The first student was in charge of methods and materials, data selection, preprocessing, model implementation, and results (tables and graphs). They had to explain the experiment, get the data ready, train the models, and assess how well they worked by looking at the visual outputs.

The Introduction, Abstract, Dataset Description, Literature Review, and References sections were written and prepared by the second student. They evaluated previous studies, established the theoretical underpinnings, and offered insights on the importance of machine learning approaches in the detection of breast cancer.

Individual contribution towards project presentation and demonstration: The first student covered the experiment process, data augmentation, and results, explaining data handling and model performance. The second student focused on theoretical concepts, dataset details, and deep learning methods, adding background and motivation. Together, they ensured a clear and well-rounded presentation.

Full Signature of Supervisor: Full signature of the student:

Full signature of the student:

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)