**DEVELOPMENT OF AN EFFICIENT PREDICTIVE MODEL USING MACHINE LEARNING ALGORITHMS FOR DISEASE CLASSIFICATION**

Supervised by Mrs. Roseline Moronkola

March 2025

In partial fulfillment of the requirement for the award of National innovation diploma In Computer Software Engineering

A Project Proposal submitted to the Department of Computer Software Engineering, SQI College of ICT, Ogbomoso

By

(230993)

Maigida Jarmai Aliyu

**INTRODUCTION**

* 1. Overview

Machine learning (ML) has become a powerful tool in cancer diagnosis by analyzing complex medical data and identifying patterns that may be hard for humans to detect. Machine learning is a subset of AI that learn from data, to improve its experience. Application of machine learning include early detection of disease and predictive modeling.

The limitations of machine learning include: Data Dependency, Human Expertise Required, and overfitting. Therefore, there is a need for deep learning to overcome these challenges.

For the course of this project, I will use a breast cancer dataset to build the model.

Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, with millions of new cases diagnosed annually (Ahmed, S. S. *et al*, 2024). The study emphasizes on the importance of women increase awareness, early detection and proactive diagnosis to reduce mortality rate (Devi, S. *et al*, 2024). Traditional diagnostic methods, such as mammography and biopsy, have notable limitations, including accuracy concerns and reliance on human expertise, which can lead to delayed or incorrect diagnoses (La Moglia, A., & Almustafa, K. M. 2025). Machine learning (ML) has emerged as a transformative tool in healthcare, offering enhanced predictive capabilities that can improve early breast cancer detection, classification, and treatment decision-making. By leveraging ML algorithms such as logistic regression, random forests, support vector machines, and deep learning models, researchers have achieved significant advancements in diagnostic accuracy, surpassing conventional methods (Zhang, B. *et al.,* 2023).

The integration of ML in medical diagnostics extends beyond breast cancer, contributing to predictive modeling for various diseases, including heart disease, diabetes, and endometriosis (Ahmed, S. S. *et al*, 2024). ML models are particularly valuable in identifying critical biomarkers and classifying tumors as benign or malignant, reducing reliance on invasive procedures and enhancing patient outcomes (La Moglia, A., & Almustafa, K. M. 2025). Explainable AI techniques, such as SHAP (Shapley Additive exPlanations) analysis, further aid in understanding model decisions, increasing trust among healthcare professionals and patients (Islam, T. *et al*., 2024). Despite its potential, challenges remain in data availability, model interpretability, and clinical integration, necessitating continued research and optimization of ML applications in oncology. As AI-driven technologies continue to evolve, their role in cancer prognosis and treatment selection will become increasingly vital in improving global healthcare systems (Zhang, B. *et al.,* 2023).

**1.2 Project Description**

This project focuses on leveraging machine learning (ML) techniques to enhance breast cancer diagnosis and prognosis. Traditional diagnostic methods, such as mammography and biopsies, often have limitations in accuracy, leading to potential misdiagnoses and delayed treatment. By implementing various ML algorithms including logistic regression, random forests, support vector machines, and deep learning this study aims to improve the accuracy and efficiency of breast cancer classification. The project utilizes a dataset containing multiple patient features, such as tumor size, patient age, and metastasis status, to train predictive models capable of distinguishing between benign and malignant tumors. Additionally, explainable AI techniques, such as SHAP analysis, will be applied to enhance model interpretability and ensure trust in automated decision-making processes.

Beyond classification, this project also explores the broader implications of ML in healthcare, particularly in treatment planning and patient prognosis. AI-driven predictive modeling can assist oncologists in identifying high-risk patients and recommending personalized treatment strategies based on tumor characteristics and historical patient data. Furthermore, integrating ML with bioinformatics can enable early detection of secondary breast cancer occurrences, ultimately improving survival rates and patient outcomes.

By addressing the challenges of traditional diagnostic methods and harnessing the power of AI, this project contributes to the advancement of precision medicine in oncology and underscores the growing role of ML in modern healthcare.

**1.3 Proposed Solution Exploration**

The project explores multiple machine learning (ML) models, including Convolutional Neural Networks (CNN), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Support Vector Classifier (SVC), and Gaussian Naïve Bayes (GNB), to enhance breast cancer diagnosis and classification. CNN is employed for image-based diagnosis, leveraging deep learning to extract spatial features from medical images, while traditional ML models like LR, RF, and GNB are used for structured data classification. XGB, known for its high accuracy and efficiency in handling complex datasets, is optimized through hyperparameter tuning, including learning rate adjustments and tree depth selection, to improve performance. SVC is incorporated for its robustness in handling high-dimensional data, ensuring precise separation between benign and malignant cases. Hyperparameter tuning techniques such as grid search, random search, and Bayesian optimization are applied to refine model performance, optimize classification accuracy, and minimize false positives/negatives. By fine-tuning these models and comparing their predictive capabilities, this study aims to identify the most effective approach for breast cancer detection, enhancing early diagnosis and improving patient outcomes.

**1.4 Importance of the Project**

This project is essential in improving breast cancer diagnosis by leveraging machine learning (ML) models to enhance accuracy, reduce diagnostic errors, and enable early detection. Traditional methods, such as mammography and biopsies, often have limitations, including human error and inconsistent interpretation, which can lead to delayed or incorrect diagnoses. By utilizing ML techniques like CNN, Random Forest, and XGBoost with hyperparameter tuning, this project optimizes predictive accuracy, helping oncologists make data-driven decisions for personalized treatment plans. Moreover, explainable AI techniques improve transparency, ensuring that ML-assisted diagnoses are interpretable and reliable for medical professionals and patients. Ultimately, this project contributes to precision medicine, enhances patient survival rates, and reduces healthcare costs by streamlining the diagnostic process and minimizing unnecessary procedures.

**1.5 Expected Results**

The project is expected to enhance breast cancer diagnosis by achieving high classification accuracy using machine learning models such as CNN, Random Forest, XGBoost, SVC, and Logistic Regression, optimized through hyperparameter tuning. By leveraging these models, the study aims to improve early detection rates, reduce false positives and negatives, and provide reliable predictions for distinguishing between benign and malignant tumors. Additionally, explainable AI techniques, such as SHAP analysis, will offer transparency into model decisions, making the results more interpretable for medical professionals. It is anticipated that deep learning models like CNN will perform best for image-based diagnosis, while ensemble methods such as XGBoost and Random Forest will excel in structured data classification. Overall, the project is expected to demonstrate the effectiveness of ML in automating cancer detection, leading to improved patient outcomes and more efficient healthcare decision-making.

**2.1 Problem Statement**

Breast cancer remains one of the leading causes of cancer-related deaths worldwide, with early and accurate detection being critical for improving patient survival rates. Traditional diagnostic methods, such as mammography and biopsies, often suffer from limitations, including human error, inconsistent interpretation, and limited accuracy, leading to delayed or incorrect diagnoses. These challenges highlight the need for more reliable, automated, and data-driven approaches to breast cancer classification. Machine learning (ML) models, including CNN, Random Forest, XGBoost, SVC, and Logistic Regression, offer promising solutions by enhancing diagnostic accuracy and reducing false positives and negatives results.

However, most exist ML models are still far away from 100% therefore is still a need to identify the most effective ML model, optimize its performance through hyperparameter tuning, and ensure its interpretability for clinical use. This project addresses these gaps by developing and evaluating ML based predictive models to improve breast cancer diagnosis, ultimately assisting healthcare professionals in making faster, more accurate, and data-driven decisions.

**3.1 Objectives**

The aim of this project is to develop an optimize machine learning (ML) models for accurate and early detection of breast cancer, reducing misdiagnosis and improving patient outcomes. Specifically, the project objectives are:

1. to implement and compare the performance of ML models, including CNN, Random Forest, XGBoost, SVC, Logistic Regression, and Gaussian Naïve Bayes, in classifying breast tumors as benign or malignant;
2. to enhance model accuracy through hyperparameter tuning and feature selection to optimize predictive performance;
3. to evaluate model performance using metrics such as specificity, sensitivity, f1-score and recall;
4. to apply explainable AI techniques, such as SHAP analysis, to improve model interpretability and ensure transparency in clinical decision-making.

By achieving these objectives, the project contributes to the advancement of AI-driven healthcare solutions, enhancing efficiency in breast cancer diagnosis and supporting early intervention strategies.

**4.1 Scope**

This project focuses on the application of machine learning (ML) techniques for breast cancer detection and classification, emphasizing the use of models such as CNN, Random Forest, XGBoost, SVC, Logistic Regression, and Gaussian Naïve Bayes. The study involves collecting and preprocessing breast cancer datasets, training ML models, and optimizing their performance through hyperparameter tuning. Additionally, explainable AI techniques, such as SHAP analysis, will be implemented to enhance model interpretability for clinical use. The scope is limited to supervised learning models for structured medical data and image-based tumor classification. While the project aims to improve diagnostic accuracy and efficiency, it does not cover real-time deployment in hospitals or the integration of patient-specific treatment recommendations. However, its findings will contribute to AI-driven healthcare advancements, potentially guiding future research and clinical applications in breast cancer diagnosis.

**5.0 Methodology**

This project follows a structured methodology to develop, optimize, and evaluate machine learning (ML) models for breast cancer detection and classification. The approach consists of the following key steps:

**5.1 Data Collection and Preprocessing**

* Obtain a breast cancer dataset from **Kaggle.com**
* Perform data cleaning, handling missing values with KNN K-nearest-Neighbor, and normalizing features to ensure consistency and accuracy.
* Apply feature selection techniques to identify the most relevant predictors for classification.

**5.2 Splits our data into training dataset and validation dataset.**

I will split the data into a **training dataset** (used to train the model) and a **validation set** (used to check performance before testing). Typically, path of the data is for training, and the remaining data is for validation. I am going to use dataset from sklearn a python library dataset.

**5.3 Model development**

* Implement multiple ML models, including Convolutional Neural Networks (CNN), Random Forest, XGBoost, Support Vector Classifier (SVC), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB).
* Train the models using labeled datasets with a focus on distinguishing between benign and malignant tumors.

**5.4 Model Evaluation**

The performance of the machine learning models used in this project will be evaluated using multiple metrics to ensure accuracy, reliability, and clinical applicability in breast cancer classification. The key evaluation metrics include:

* **Accuracy** – Measures the overall correctness of the model by calculating the ratio of correctly classified instances to the total instances. While accuracy is useful, it may be misleading in cases of imbalanced datasets.
* **Precision** – Evaluates how many of the predicted positive cases (malignant tumors) are actually correct. This is crucial in minimizing false positives.​
* **Recall (Sensitivity)** – Measures how well the model identifies actual positive cases, ensuring that malignant cases are not missed.
* **F1-Score** – Provides a balance between precision and recall, especially useful when the dataset is imbalanced.
* **Specificity** – Measures how well the model identifies actual negative cases (benign tumors), ensuring that false positives are minimized. Each machine learning model, including CNN, Random Forest, XGBoost, SVC, Logistic Regression, and Gaussian Naïve Bayes, will be tested using these metrics. A confusion matrix will also be used to visualize the model’s classification performance. Cross-validation techniques will be applied to ensure robustness and prevent overfitting. The best-performing model will be selected based on a combination of high accuracy, balanced precision-recall, and interpretability, ensuring it is suitable for real-world breast cancer diagnosis.

**5.5 Model explanation.**

The model explanation is based onOncology Predictions Using SHAP (Shapley Additive Explanations).

**Oncology** is the branch of medicine that deals with the prevention, diagnosis, and treatment of cancer.

**SHAP** is a method based on game theory that explains the output of machine learning models by assigning **feature importance scores**. It helps us understand **how much each feature contributes** to a model's prediction

**5.6 Model validation**

To evaluate the effectiveness of the trained machine learning models, a separate testing dataset will be used to measure their real-world performance. After training the models on a labeled dataset, the testing phase will involve feeding unseen data into the models to assess their ability to classify breast tumors as benign or malignant. The testing dataset will be preprocessed similarly to the training dataset to ensure consistency in feature scaling and normalization.

During testing, the model’s predictions will be compared against the actual labels in the dataset. Performance metrics such as accuracy, precision, recall, F1-score, and specificity will be calculated to determine how well the model generalizes to new data. Additionally, confusion matrices will be used to analyze false positives and false negatives, which are critical in medical diagnostics. Cross-validation may also be applied to further validate model robustness. The final evaluation will help determine the most effective model for breast cancer detection, ensuring it is reliable and accurate for potential clinical applications.

5.7 **Comparison and Selection of the Best Model**

* Compare the performance of all implemented models to identify the most effective approach.
* Determine the best-performing model based on diagnostic accuracy, interpretability, and computational efficiency.
  + 1. **Report and Recommendations**
* Document findings, including model performance insights and potential clinical applications.
* Provide recommendations for integrating ML models into healthcare systems for enhanced breast cancer diagnosis.
  1. **Model deployment using Django API**

This project uses machine learning to predict whether a tumor is benign (non-cancerous) ormalignant (cancerous) based on medical features. To make this prediction accessible to users, I deployed the model using a **Django API**

How the it works:

1. **Data Processing & Model Training:**

* The model was trained using historical breast cancer data.
* Important medical features (like tumor size and texture) were used to make predictions.
* After training, the model was saved for future use.

1. **Django API Deployment:**

* The trained model was integrated into a Django API.
* Users can send their sample dataset to the API.
* The API processes the input and runs the model to predict whether the tumor is Benign or Malignant.

1. **How Users Get Predictions:**

* A user submits data via the API (e.g., a web form).
* The API returns a response to user

This methodology ensures a systematic and data-driven approach to improving breast cancer classification, contributing to more accurate and efficient diagnostic tools in medical practice.

**5.8 Significance**

The project significant is to advancing breast cancer diagnosis by leveraging machine learning (ML) models to improve accuracy, efficiency, and early detection. By integrating ML techniques like CNN, Random Forest, XGBoost, and SVC, this study enhances predictive capabilities, reduces misclassification rates, and minimizes false positives and negatives, leading to more reliable diagnoses.

Furthermore, this project contributes to precision medicine by enabling data-driven decision-making in oncology. Explainable AI techniques, such as SHAP analysis, improve model transparency, ensuring medical professionals can trust and interpret ML-driven results. By optimizing ML models through hyperparameter tuning, the study ensures robust and efficient classification, ultimately supporting oncologists in making faster and more accurate clinical decisions. This research not only enhances breast cancer screening but also sets a foundation for integrating AI-based solutions into broader healthcare applications, improving patient outcomes and reducing healthcare costs.

**Reference**

1. Ahmed, S. S., Srivastava, Y., & Khan, M. G. (2024). Prediction and Diagnosis of Breast Cancer Using Machine Learning Algorithms. *Asian Journal of Research in Medical and Pharmaceutical Sciences*, *13*(3), 54–60. <https://doi.org/10.9734/ajrimps/2024/v13i3261>
2. **Almarri, B., Gupta, G., Kumar, R., et al.** (2024). The BCPM method: Decoding breast cancer with machine learning. BMC Medical Imaging, 24, 248. <https://doi.org/10.1186/s12880-024-01402-5>
3. Devi, S., Kaul Ghanekar, R., Pande, J. A., Dumbre, D., Chavan, R., & Gupta, H. (2024). Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models. *Asian Pacific journal of cancer prevention: APJCP*, *25*(3), 1077–1085. <https://doi.org/10.31557/APJCP.2024.25.3.1077>
4. **Goel, A., Chauhan, A., Pal, D., & Singh, R.** (2024). Analysis of machine learning algorithms for prediction and classification of breast cancer. International Journal of Computer Applications, 186(20), 43–48. <https://doi.org/10.5120/ijca2024923632>
5. Minnoor, M., & Baths, V. (2023). Diagnosis of breast cancer using random forests. *Procedia Computer Science, 218*, 429–437. <https://doi.org/10.1016/j.procs.2023.01.025>
6. Islam, T., Sheakh, M. A., Tahosin, M. S., Hena, M. H., Akash, S., Bin Jardan, Y. A., FentahunWondmie, G., Nafidi, H. A., & Bourhia, M. (2024). Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Scientific reports*, *14*(1), 8487. <https://doi.org/10.1038/s41598-024-57740-5>
7. **Jagetiya, A., & Dadhech, P.** (2024). Optimizing breast cancer prognosis with machine learning for enhanced clinical decision-making. Bio-Algorithms and Med-Systems, 20(1), 37–48. <https://doi.org/10.5604/01.3001.0054.8929>
8. Jakkaladiki SP, Maly F. 2023. An efficient transfer learning based cross model classification (TLBCM) technique for the prediction of breast cancer. *PeerJ Computer Science* 9: e1281 <https://doi.org/10.7717/peerj-cs.1281>
9. Kawina, I., K., A., & Marapelli, B. (2024). Deep learning and machine learning approach to breast cancer classification with random search hyperparameter tuning. *International Journal of Intelligent Systems and Applications in Engineering, 12*(16s), 264–275. Retrieved from <https://www.ijisae.org/index.php/IJISAE/article/view/4818>
10. **Kumar,** **B. N. R., Gowda, B. J. N. C., A., H. N., V., B., Sujitha, B. B., & Ramani, D. R.** (2024). An efficient breast cancer detection using machine learning classification models. International Journal of Online and Biomedical Engineering (iJOE, 20(13), 24–40. <https://doi.org/10.3991/ijoe.v20i13.50289>
11. La Moglia, A., & Almustafa, K. M. (2025). Breast cancer prediction using machine learning classification algorithms. Intelligence-Based Medicine, 11, 100193. <https://doi.org/10.1016/j.ibmed.2024.100193>
12. Madeswaran, T. S., Kavuru, A. K., Theagarajan, P., Al Hadhrami, N., Al Foori, M., & Rambabu, O. (2023). An intelligent system for predicting the breast cancer threat using health data registry and awareness: A review. *European Journal of Engineering and Technology Research, 8*(3), 17–22. <https://doi.org/10.24018/ejeng.2023.8.3.3012>
13. Manalı, D.; Demirel, H.; Eleyan, A. (2024). Deep Learning Based Breast Cancer Detection Using Decision Fusion. *Computers* 2024, 13, 294. <https://doi.org/10.3390/computers13110294>
14. **Mahmood, D. A., & Aminfar, S. A.** (2024). Efficient machine learning and deep learning techniques for detection of breast cancer tumor. BioMed Target Journal, 2(1), 1–13. <https://doi.org/10.59786/bmtj.211>
15. **Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Ait Abdelouhahid, R., & Debauche, O.** (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487–492. <https://doi.org/10.1016/j.procs.2021.07.062>
16. Nazari, E., Naderi, H., Tabadkani, M., et al. (2023). Breast cancer prediction using different machine learning methods applying multi factors. *Journal of Cancer Research and Clinical Oncology, 149*, 17133–17146. <https://doi.org/10.1007/s00432-023-05388-5>
17. Okebule, T., Adeyemo, O. A., Oguntimilehin, A., Obamiyi, S. E.., & Badeji-Ajisafe, B. (2023). Machine Learning Techniques for Breast Cancer Prediction: A Concise Review. *ABUAD International Journal of Natural and Applied Sciences*, *3*(2), 76–86. <https://doi.org/10.53982/aijnas.2023.0302.11-j>
18. Clift, A. K., Dodwell, D., Lord, S., Petrou, S., Brady, M., Collins, G. S., et al. (2023). Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: Cohort study. *BMJ*, *381*, e073800. <https://doi.org/10.1136/bmj-2022-073800>
19. Raha, A. D., Dihan, F. J., Gain, M., Murad, S. A., Adhikary, A., Hossain, M. B., Hassan, M. M., Al-Shehari, T., Alsadhan, N. A., Kadrie, M., & Bairagi, A. K. (2024). Modeling and predictive analytics of breast cancer using ensemble learning techniques: An explainable artificial intelligence approach. *Computers, Materials & Continua, 81*(3), 4033–4048. <http://www.techscience.com/cmc/v81n3/59042>
20. **Rahman, R., Saha, D., Dkhar, W., et al.** (2025). Development of a machine learning predictive model for early detection of breast cancer [Version 1; peer review: 1 approved with reservations, 1 not approved]. F1000Research, 14, 164. <https://doi.org/10.12688/f1000research.161073.1>
21. Raza, M. A., Khattak, A. M., Abbas, W., & Asghar, M. Z. (2024). Efficient diagnoses of breast cancer disease using deep learning technique. *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence (ICCAI '24)*, 136–143. Association for Computing Machinery. <https://doi.org/10.1145/3669754.3669775>
22. **Selvaraj, S., Thangavel, S., Prabhakaran, M., & Sathish, T.** (2024). Impressive predictive model for breast cancer based on machine learning. EAI Endorsed Transactions on Pervasive Health and Technology, 10. <https://doi.org/10.4108/eetpht.10.5246>
23. **Song, X., Chu, J., Guo, Z., Wei, Q., Wang, Q., Hu, W., Wang, L., Zhao, W., Zheng, H., Lv, X., & Zhou, J.** (2024). Prognostic prediction of breast cancer patients using machine learning models: A retrospective analysis. Gland Surgery, 13(9). <https://gs.amegroups.org/article/view/129247>
24. **Sultana, J.** (2024). A quick review on breast cancer prediction using machine learning. International Journal of Engineering Technology & Informatics, 5(1), 1–3. doi:10.51626/ijeti.2024.05.00068
25. Sun, Y., Xia, X., & Liu, X. (2024). Predictive modeling of breast cancer-related lymphedema using machine learning algorithms. *Gland Surgery, 13*(12), 2243–2252. <https://doi.org/10.21037/gs-24-252>
26. Wang, S. (2024). Design and implementation of breast cancer prediction system based on machine learning. *Theoretical and Natural Science, 32*, 238-247. <https://doi.org/10.54254/2753-8818/32/20240871>
27. Yavuz, Ö. Ç., Calp, M. H., & Erkengel, H. C. (2023). Prediction of breast cancer using machine learning algorithms on different datasets. *Ingeniería Solidaria, 19*(1), 1–32. <https://doi.org/10.16925/2357-6014.2023.01.08>
28. Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. *Journal of multidisciplinary healthcare*, *16*, 1779–1791. <https://doi.org/10.2147/JMDH.S410301>
29. **Zhou, S., Hu, C., Wei, S., & Yan, X.** (2024). Breast cancer prediction based on multiple machine learning algorithms. Technology in Cancer Research & Treatment, 23. <https://doi.org/10.1177/15330338241234791>
30. **Zuo, D., Yang, L., Jin, Y., et al.** (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Medical Informatics and Decision Making, 23, 276. <https://doi.org/10.1186/s12911-023-02377-z>