INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 15, Issue 2, (2025) E-ISSN: 2222-6990

A Comparative Study of Features Selections in Breast Cancer Classification Using Machine Learning Algorithms

Muhammad Bilal Din[1], Paridah Daud[1*], Noor Azma Ismail[1], Noor Lees Ismail[1], Farhad Nadi[1]

[1]School of Information Technology, UNITAR International University, Selangor, Malaysia

Corresponding Author Email: paridah69@unitar.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v15-i2/24705 DOI:10.6007/IJARBSS/v15-i2/24705

Published Date: 15 February 2025

Abstract

Breast cancer remains a leading cause of mortality among women worldwide, with early detection being critical for reducing mortality rates. This study aimed to evaluate different ML algorithms for breast cancer prediction and to identify key features contributing to accurate classification. By utilizing various feature selection techniques and analyzing their impact on model performance, the study provides critical insights into optimizing machine learning models for breast cancer diagnostics. Using a dataset of mammographic images, this study evaluates various feature selection approaches to distinguish between benign and malignant cases. The selected features are analyzed using machine learning algorithms Naïve Bayes, logistic Functions, Sequential Minimal Optimization (SMO), and Decision Trees. Results demonstrate that effective feature selection enhances classification accuracy, with SMO and naïve bayes algorithms achieving the highest performance accuracy 96.996 with full features provided. This study also explores the influence of differences features selection technique to compare the performance of machine learning methods for breast cancer detection.

Keywords: Machine Learning Algorithms, Mammographic Image Analysis, Feature Selection, Supervised algorithms, Decision Tree

Introduction

Breast cancer is one of the most prevalent and life-threatening diseases affecting women globally, particularly those in middle age. Despite advancements in medical science, the mortality rate associated with breast cancer remains alarmingly high, primarily due to delayed diagnosis and limited access to accurate diagnostic tools. Mammography, the most widely used screening method, is effective but often constrained by manual interpretation and variability in expertise. To address these limitations, this research leverages advancements in machine learning (ML) and image processing to improve diagnostic accuracy and reliability. By automating the analysis of mammographic images and integrating advanced ML

algorithms, this study aims to enhance early detection, reduce diagnostic errors, and improve patient outcomes.

Problem Statement

Breast cancer remains one of the leading causes of mortality among women worldwide, with early detection being critical to improving survival rates. Traditional diagnostic methods, including mammography and biopsy, are often subjective, time-consuming, and prone to human error. The variability in radiologists' interpretations can lead to misdiagnosis, resulting in either unnecessary treatment or delayed interventions. Machine learning (ML) algorithms have emerged as a promising solution for enhancing breast cancer diagnostics by automating classification processes and improving accuracy. However, an essential challenge in ML-based breast cancer classification is identifying the most relevant features that contribute to accurate predictions. With an overwhelming number of features extracted from mammographic images, the inclusion of irrelevant or redundant features can reduce model efficiency and lead to computational complexities. This study aims to address this gap by comparing different feature selection techniques and their impact on ML algorithms in breast cancer classification. By evaluating the performance of various classifiers with different feature subsets, the research seeks to optimize ML-based diagnostic systems, ultimately improving detection accuracy and assisting healthcare professionals in early-stage cancer identification.

Research Objectives

- To evaluate performance of different machine learning algorithms for predicting Breast Cancer
- To provide insights into the key features contributing to Breast Cancer prediction.

This research addresses a critical gap in breast cancer diagnostics by proposing an ML-based system designed to improve the accuracy and reliability of mammographic image analysis. By integrating advanced feature extraction methods and supervised learning algorithms, the study provides a scalable solution for early detection and classification of breast cancer. The outcomes highlight the potential of ML in reducing diagnostic errors, empowering healthcare professionals, and ultimately saving lives.

Literiture Review

The selection of appropriate methods for classification, feature selection, and segmentation in medical imaging and bioinformatics plays a crucial role in advancing healthcare technology. The studies reviewed span across various techniques used to improve data handling in the classification of diseases like breast cancer, identifying regions of interest (ROI) in mammograms, and segmenting images for improved diagnostic analysis(Ortiz et al., 2024). Siyabend Turgut et al. employed microarray data to classify breast cancer patients using eight machine learning algorithms such as SVM, KNN, MLP, and Decision Trees. The study highlighted the significance of feature selection methods like Recursive Feature Elimination (RFE) and Randomized Logistic Regression (RLR) in improving classification accuracy. Notably, SVM showed the best results post-feature selection, showcasing the critical role of selecting relevant features for effective patient classification(Turgut et al., 2018).

Varalatchoumy et al. presented a comparative study of four innovative approaches for detecting regions of interest (ROI) in mammograms. These approaches incorporated methods like histogram equalization, morphological operations, and advanced segmentation algorithms. The fourth approach, combining adaptive histogram equalization and novel algorithms, proved the most efficient and reliable for radiologists, especially for real-time hospital data(Ravishankar & Varalatchoumy, 2017). Ammu P. K. et al. reviewed various feature selection methods for DNA microarray data, including Biogeography-Based Optimization (BBO) and Particle Swarm Optimization (PSO). They emphasized the importance of removing redundant genes for better representation of target classes. The study also explored hybrid methods integrating filtering and genetic algorithms to enhance gene selection accuracy, critical for bioinformatics applications(Pk & V, 2013).

Lan Li et al. introduced a Fuzzy Level Set algorithm for automated medical image segmentation. This method integrated spatial fuzzy clustering to initialize segmentation and optimized parameters for robust performance. Enhanced with locally regularized evolution, the algorithm showed promising results across various medical imaging modalities, advancing automation in medical diagnostics (Li et al., 2011). Yao et al. (2022), explored machine learning techniques applied to fine needle aspiration (FNA) biopsy data, emphasizing artificial neural networks (ANNs) for classifying malignant and benign cases. The use of k-fold cross-validation ensured model robustness, highlighting the role of ML algorithms in enhancing diagnostic accuracy for breast cancer(Shafique et al., 2023).

Shoaib Farooq and Ilyas analyzed environmental influences on breast cancer using machine learning models. By examining factors like air quality and socioeconomic status, their study demonstrated the capacity of ML to identify complex patterns in vast datasets. This integration of environmental data emphasizes the potential for early detection and personalized treatment strategies(Farooq & Ilyas, 2023). Liu et al. developed a model integrating clinical, molecular, and pathology data to predict breast cancer recurrence and metastasis risk. The study highlighted the superiority of multifaceted ML models over traditional methods, showcasing the importance of comprehensive data integration for more accurate prognostic assessments (Yao et al., 2022). Li et al. discussed the impact of data quality and completeness on ML model performance. They emphasized rigorous data collection and validation processes to mitigate biases, addressing a critical challenge in developing reliable ML models for breast cancer diagnostics and treatment(Budach et al., 2022).

Li et al. also reviewed the application of ML in immunotherapy for breast cancer. They noted how ML techniques enhance diagnosis, grading, and prognosis by leveraging complex datasets. This underscores the transformative role of ML in oncology, particularly in optimizing clinical decision-making(Cruz & Wishart, 2007). Dhiman et al. identified common flaws in oncology ML models, such as small sample sizes and inadequate preprocessing. Their critique underlined the need for rigorous methodological standards to improve the predictive reliability of ML applications in breast cancer survival analysis. This call for better practices aligns with the broader goal of advancing ML's role in healthcare(Dhiman et al., 2022). The integration of histopathological images with clinical and genomic data is a promising direction for improving breast cancer recurrence and metastasis prediction. By using the ICSDA model, it is possible to predict recurrence and metastasis more accurately and cost-effectively,

ultimately improving the prognostic capabilities for clinicians and offering better outcomes for patients. The study emphasizes the potential of combining multi-modal data for predictive modelling in cancer diagnosis, reducing reliance on high-cost genomic sequencing while leveraging widely available histopathological images (Turgut et al., 2018).

Numerous AI techniques and products have been developed to support cancer treatment and prevention in health facilities and communities. Systematic reviews of these efforts consistently show that AI-assisted interventions generally outperform conventional methods. However, challenges remain, such as the need for rigorous evaluation of predictive models, as noted by Lisboa et al., and the importance of understanding the complex structure of datasets and individual factors. Ray et al. pointed out the lack of cloud computing and long-range communication in wearable systems for cancer detection, emphasizing the potential of AI and machine learning in improving these technologies. While AI-based imaging applications for breast cancer diagnosis have demonstrated superior performance, most studies lack high-level evidence, underscoring the need for further clinical research and health technology assessments (Tran et al., 2019).

Using a breast cancer dataset from the University of California Irvine (UCI) machine learning repository including 569 records and 32 attributes, a supervised machine learning method was applied to identify whether a tumour is benign or malignant. Using patient traits and tumour parameters like radius, texture, perimeter, and symmetry, the k-Nearest Neighbour (k-NN) algorithm was applied. To develop the k-NN model, the data was separated into a training set including 469 records and a testing set comprising the remaining records for simulating new patients. Features were rescaled using data normalising into a conventional range. Using the starting k-value of 21—about the square root of the size of the training dataset—alternative k-values (1, 5, 11, 15, 21, 27) were investigated to maximise model performance. R's "class" package was used in the analysis (v3.6.2) (Ayde et al., 2023).

The biggest Breast cancer now kills more women than heart disease. Genetic factors are important in breast cancer growth, but recent research suggests environmental variables are too. This study reviews environmental factors that may affect breast cancer risk, incidence, and outcomes. The study examines how lifestyle choices including diet, exercise, and alcohol use affect hormonal imbalances and inflammation, two key breast cancer risk factors. It also discusses pesticides, endocrine-disrupting chemicals (EDCs), and industrial pollutants, which interfere with hormone signalling and DNA damage and increase breast cancer risk. Machine learning algorithms make predictions. Logistic Regression, Random Forest, KNN, SVC, additional tree classifier. Model evaluation metrics included confusion matrix correlation coefficient, F1-score, Precision, Recall, and ROC curve. The classifier with the highest accuracy is Random Forest with 0.91% accuracy and Logistic Regression ROC curve 0.901%. The many machine learning algorithms used in this research were accurate, indicating that they could replace predicting methods in breast cancer survival analysis, particularly in Asia(Farooq & Ilyas, 2023).

Deep learning could be used to pre-screen cancer by analysing demographic and anthropometric data, biological markers from routine blood tests, and relative risks from meta-analysis and international databases. We used feature selection algorithms to find the top cancer pre-screening predictors in 116 women, comprising 52 healthy and 64 breast

cancer patients. We used the best predictors to compare deep learning to classical machine learning techniques in k-fold Monte Carlo cross-validation tests. Our results show that a deep learning model with an input-layer architecture fine-tuned via feature selection can identify cancer patients from non-cancer patients. Deep learning predicts with the lowest uncertainty compared to machine learning. The results show that deep learning algorithms can supplement imagery-based cancer pre-screening with radiation-free, non-invasive, and cheap methods. Deep learning algorithms in cancer pre-screening can identify people who need imaging-based screening, motivate self-examination, and reduce the psychological externalities of false positives. Deep learning algorithms for screening and pre-screening will detect cancer early, saving healthcare and societal costs(Martinez & Dongen, 2023).

Wearable sensors are gaining popularity in cancer care, but their full usefulness depends on their ability to reliably translate raw outputs into AI/ML-ready data. This review indicated that researchers are employing diverse preprocessing methods to address this difficulty, but with no conventional best practices. We found a need for uniform data quality and preprocessing methods for wearable sensor data to support cancer research and diverse patient populations. Given the variety of preprocessing methods in the literature, a framework to help researchers and clinicians prepare wearable sensor data for AI/ML applications is needed. In the scoping review and our study, we provide a general paradigm for preprocessing wearable sensor data that is extensible to disease contexts outside cancer treatment(Ortiz et al., 2024).

Technology reduces breast cancer mortality in this vital subject. Many ML techniques have been developed for medical dataset analysis. Breast cancer data must be accurately and effectively classified for medical diagnosis. Many approaches have been developed to classify breast cancer data, but accuracy remains a challenge. We proposed a breast cancer data categorisation model to solve this issue. This paper used WDBC dataset and six ML classification methods: DT, KNN, SVM, RF, NB, and LR with ensemble methods. All techniques are compared, and Decision Tree classifier with criterion gini index had the highest accuracy: 97% and AUC: 0.996 for LR classifier, while XGBoost had the highest accuracy: 97% and AUC 0.99 for ensemble techniques. The proposed ensemble learning method may help cancer doctors identify cancer(More, 2022).

Research Methodology

The methodology outlines the research design, data collection methods, data processing steps, analytical procedures, and tools utilized to study the factors influencing breast cancer patient survival.

Research Design

This research study was employed a retrospective observational design. The dataset consists of historical records of breast cancer patients, with features such as Cell Size Uniformity, Cell Shape Uniformity, Bland Chromatin, Clump_Thickness, Normal_Nucleoli, Marginal_Adhesion, and Single_Epi_Cell_Size.

Data Collection

The dataset used for this research is a retrospective collection of breast cancer patient data, including key variables like age, tumor size, lymph node involvement, and survival outcomes.

This data provides a basis for understanding factors that influence survival and treatment outcomes.

Experimental Setup Framework

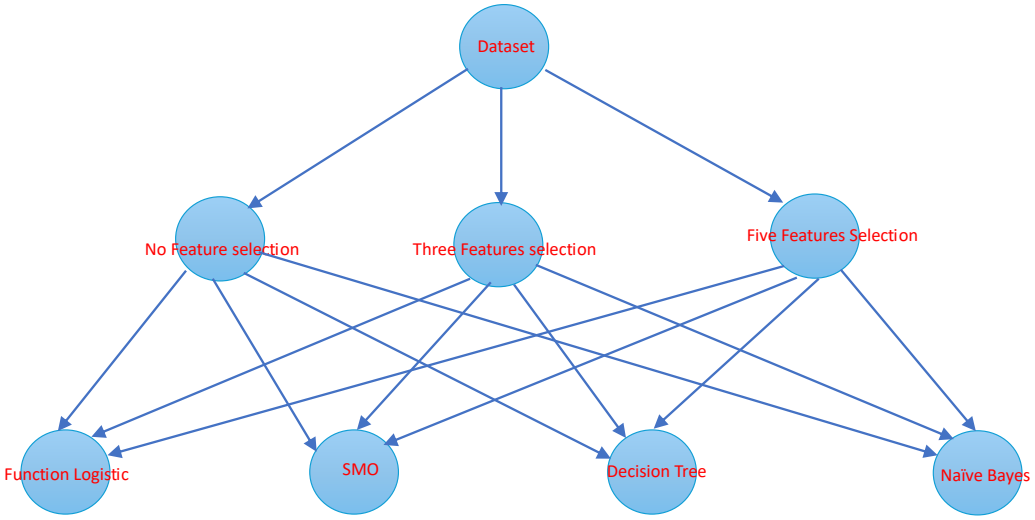

Figure 1: Comparison between algorithms based on features selection from dataset

The analysis or model. Once the relevant attributes are identified, the next step is to set the target role, which defines the outcome or dependent variable that the analysis aims to predict or explain. Thus, feature extraction is performed, where raw data is transformed into more useful and meaningful features, potentially enhancing the performance of any subsequent analysis or machine learning models. The process then ends, ready for the analysis or model training.

Feature Selection

Feature selection was conducted to improve model accuracy and efficiency. Key selected features included:

- Cell_Shape_Uniformity
- Cell_Size_Uniformity
- Bare_Nuclei
- Bland_Chromatin
- Clump_Thickness
- Normal_Nucleoli
- Marginal_Adhesion
- Single_Epi_Cell_Size

Predictive Modelling

Logistic Regression: Construct a binary logistic regression model to predict survival probability based on significant predictors identified from prior analysis.

Tree-Based Models (Decision Trees and Random Forest): Build tree-based models to capture non-linear relationships and interactions among features.

Model Evaluation: Evaluate models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Implement k-fold cross-validation to validate the robustness of the model.

Modelling and Analysis

Model Selection Multiple classifiers were trained and evaluated in WEKA to determine the best model for survival prediction. The following ML algorithms models were used are:

- Decision Tree: Chosen for its interpretability and ability to manage non-linear relationships.
- SMO: Selected for its robustness and accuracy, especially when handling feature interactions.
- Naïve Bayes: Included for baseline comparison due to its simplicity and quick computation.
- Logistic Functions: Utilized to assess linear relationships between survival and selected features.

Model Training and Evaluation Each model was trained on 70% of the dataset, with 30% reserved for testing. WEKA's Cross-validation (10-fold) to validate model performance, ensuring reliability and reducing overfitting risks. Performance Metrics Model performance was evaluated using metrics such as:

- Accuracy: Overall correctness of the model in predicting survival status.
- Precision and Recall: Assessed the model's ability to correctly classify Alive vs. Deceased.
- ROC-AUC: Used to evaluate the area under the Receiver Operating Characteristic curve, especially for models predicting binary outcomes.

Result and Discussion

An implementation and analysis overview typically involves a summary of executing a project or system and evaluating its effectiveness. Implementation focuses on setting up, configuring, or launching the project to achieve desired outcomes, ensuring all components function as intended. Analysis, on the other hand, examines the results, measuring performance, identifying issues, and assessing the impact. Together, these steps help refine processes, improve system efficiency, and provide insights for future improvements, contributing to overall project success.

Comparison algorithms Models analysis based on No Feature Selection Table 1 presented comparison the performance metrics of different machine learning models: Logistic, Decision Tree, SMO, and Naive Bayes for No Feature Selection.

Table 1
Comparison Algorithms models based on No Feature Selection

Model	Accuracy (%)	Precision	Recall	F-Measure	ROC Area
Logistic	96.567	0.966	0.966	0.966	0.993
Decision Tree	95.280	0.953	0.953	0.953	0.987
SMO	96.996	0.970	0.970	0.970	0.968
Naive Bayes	95.990	0.962	0.960	0.960	0.986

SMO outperformed other models in terms of accuracy, precision, recall, and F-Measure, indicating it is the best overall performer. Logistic regression shows high ROC Area (0.993), suggesting good classification performance. Decision Tree had the lowest performance across most metrics but still achieved decent results. Naive Bayes had good accuracy, but slightly lower recall and F-measure compared to SMO and Logistic. This visual comparison helps in determining the most effective model for the task, with SMO standing out as the best choice.

Comparison algorithms Models analysis based on Three Features Selection
Based on the nature of the dataset, Table 2 presented comparison the performance metrics of different machine learning models: Logistic, Decision Tree, SMO, and Naive Bayes for three Features Selection.

- Marginal_Adhesion
- Cell_Size_Uniformity
- Clump Thickness

Table 2
Comparison Algorithms models based on Three Feature Selection

Model	Accuracy (%)	Precision	Recall	F-Measure	ROC Area
Decision Tree	94.85	0.94	0.94	0.94	0.96
Logistic	94.71	0.94	0.94	0.94	0.99
Naive Bayes	94.56	0.94	0.94	0.94	0.99
SMO	94.85	0.94	0.94	0.94	0.96

All models demonstrate similar performance, with Decision Tree and SMO slightly leading in Accuracy. Logistic Regression and Naive Bayes excel in ROC Area, indicating strong ability to distinguish between classes. The choice of the best model may depend on the specific application requirements, as differences in performance metrics are minimal

Comparison algorithms Models analysis based on Five Features Selection
Based on the nature of the dataset, Table 3 presented comparison the performance metrics of different machine learning models: Logistic, Decision Tree, SMO, and Naive Bayes for 5 Features Selection attributes

- Single EPI Cell Size
- Cell Shape Uniformity
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli

Table 3
Comparison Algorithms Models performance based on Five Feature Selection

Model	Accuracy (%)	Precision	Recall	F-Measure	ROC Area
Decision Tree	94.42	0.94	0.94	0.94	0.95
Logistic	95.42	0.94	0.95	0.95	0.99
Naïve Bayes	95.71	0.95	0.94	0.95	0.90
SMO	95.99	0.96	0.96	0.96	0.96

SMO emerges as the top-performing model with the highest accuracy and strong metrics in Precision, Recall, and F-Measure. Logistic Regression and Naive Bayes show slightly lower accuracy but excellent ROC Area (~0.99), suggesting strong classification capabilities. Decision Tree performs well but has the lowest Accuracy and ROC Area among the models. Overall, all models demonstrate robust performance, but SMO is the most accurate, while Logistic and Naive Bayes excel in distinguishing between cl asses with their high ROC Area.

Conclusion

This study aimed to evaluate different ML algorithms for breast cancer prediction and to identify key features contributing to accurate classification. By utilizing various feature selection techniques and analyzing their impact on model performance, the study provides critical insights into optimizing machine learning models for breast cancer diagnostics.
Table 4 presented the final comparative analysis for all ML algorithms with differences of feature selection.

Table 4
Comparative Analysis Table for Cross-Validation 10 Folds

Feature Selection	Algorithms	Accuracy (%)	Precision	Recall	F-Measure	ROC Area
No Feature Selection	Function Logistic	96.567	0.966	0.966	0.966	0.993
	SMO	96.996	0.953	0.970	0.970	0.968
	Decision Tree	95.280	0.953	0.953	0.953	0.987
	Naïve Bayes	96.996	0.970	0.960	0.960	0.986
Three Features Selection	Function Logistic	94.710	0.940	0.940	0.940	0.990
	SMO	94.850	0.940	0.940	0.940	0.990
	Decision Tree	94.850	0.940	0.940	0.940	0.990
	Naïve Bayes	94.560	0.940	0.940	0.940	0.990
Five Features Selection	Function Logistic	95.420	0.940	0.950	0.950	0.950
	SMO	95.990	0.960	0.960	0.960	0.960
	Decision Tree	94.420	0.940	0.940	0.940	0.940
	Naïve Bayes	95.720	0.960	0.940	0.940	0.940

The results indicate that feature selection significantly influences classification accuracy, precision, recall, F-measure, and ROC area. Without feature selection, models such as SMO and Naïve Bayes achieved the highest accuracy (96.996%), suggesting that a full feature set provides better classification power. However, applying three or five feature selections resulted in performance degradation across all models, with three-feature selection yielding the lowest accuracy range (94.560% - 94.850%). This underscores the importance of selecting an optimal feature subset to balance model complexity and predictive power.

Among the evaluated algorithms, SMO and Naïve Bayes demonstrated stable performance across different feature selection settings, indicating their robustness in breast cancer classification tasks. Decision Tree, however, showed the highest sensitivity to feature selection, with a significant drop in accuracy when fewer features were used. This suggests that decision trees rely on a broader range of features for accurate classification.

The study successfully demonstrates that machine learning algorithms can enhance the accuracy and reliability of breast cancer classification. However, careful feature selection is crucial in optimizing performance while minimizing computational overhead.
Key takeaways from the study include:

- SMO and Naïve Bayes are the most effective models for breast cancer prediction, achieving the highest accuracy with and without feature selection.
- Reducing the number of features impacts model performance, with three-feature selection leading to a significant decrease in predictive accuracy.
- Feature selection strategies should be tailored based on the specific application, as excessive feature reduction can lead to loss of critical information.

This research contributes to the growing body of knowledge on ML-based breast cancer diagnostics by emphasizing the importance of feature selection. The findings highlight the potential of machine learning in improving early detection, reducing diagnostic errors, and assisting healthcare professionals in making informed decisions. Future work could focus on refining feature selection methodologies and integrating deep learning techniques to enhance classification performance further.

Future Research Directions

Future research can focus on integrating deep learning techniques such as convolutional neural networks (CNNs) to further enhance breast cancer classification accuracy. Additionally, incorporating hybrid feature selection methods that combine statistical and ML-based approaches could optimize feature extraction. Investigating real-time implementation in clinical settings and enhancing interpretability through explainable AI techniques would also be beneficial. Lastly, research into personalized treatment recommendations based on ML predictions could significantly impact patient care.

Significance to Society

This study contributes to society by improving breast cancer diagnostics, potentially leading to earlier detection and better patient outcomes. By reducing diagnostic errors, ML-based systems can help healthcare professionals make more accurate and efficient decisions. The research also promotes cost-effective healthcare solutions by minimizing unnecessary biopsies and medical interventions. Ultimately, the integration of ML in breast cancer classification can enhance public health strategies, reduce mortality rates, and improve the quality of life for individuals at risk of breast cancer.

Theoretical and Contextual Contribution

This study contributes to both theoretical and practical domains in the field of breast cancer diagnostics using machine learning (ML). Theoretically, it extends the existing knowledge base by demonstrating the impact of feature selection techniques on classification accuracy. The findings highlight the importance of selecting relevant features to optimize ML models for

medical image analysis, providing a structured approach for feature selection in healthcare AI applications. This research also emphasizes the significance of supervised learning algorithms in enhancing the precision of breast cancer classification, which adds to the growing body of literature on AI-driven medical diagnostics.

From a contextual perspective, this study plays a crucial role in improving diagnostic accuracy and efficiency within healthcare settings. By evaluating multiple ML models, it provides valuable insights into their comparative performance, enabling healthcare practitioners and researchers to make informed decisions about algorithm selection. The research outcomes can be particularly beneficial in low-resource medical facilities, where ML-based automated screening tools can assist radiologists in early detection, reducing reliance on subjective assessments. Additionally, by advocating for optimal feature selection methods, this study paves the way for more efficient and cost-effective computational models that can be integrated into clinical workflows, ultimately enhancing patient care and survival rates.

Acknowledgement: The publication of this paper was supported by UNITAR International University, Malaysia

References

Ammu, P. K., & Preeja, V. (2013). Review on feature selection techniques of DNA microarray data. *International Journal of Computer Applications*, *61*(12).

Ayde, C. C. N., Magda, M. M. I., Epifania, C. P. S., & Fred, T.-C. (2023). Prediction of breast cancer with 98% accuracy. *arXiv*. https://doi.org/10.48550/arXiv.2307.07571

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022). The effects of data quality on machine learning performance (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2207.14529

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

Dhiman, P., Ma, J., Andaur Navarro, C. L., Speich, B., Bullock, G., Damen, J. A. A., Hooft, L., Kirtley, S., Riley, R. D., Van Calster, B., Moons, K. G. M., & Collins, G. S. (2022). Methodological conduct of prognostic prediction models developed using machine learning in oncology: A systematic review. *BMC Medical Research Methodology, 22*(1), 101.

Farooq, M. S., & Ilyas, M. (2023). Predicting environment effects on breast cancer by implementing machine learning. *arXiv*. https://doi.org/10.48550/arXiv.2309.14397

Li, B. N., Chui, C. K., Chang, S., & Ong, S. H. (2011). Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Computers in biology and medicine*, *41*(1), 1-10.

Lisboa, P. J., & Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, *19*(4), 408-415.

Martinez, R. G., & van Dongen, D.-M. (2023). Pre-screening breast cancer with machine learning and deep learning. *arXiv*. https://doi.org/10.48550/arXiv.2302.02406

More, A. (2022). Breast cancer prediction using classification techniques of machine learning. *International Journal for Research in Applied Science and Engineering Technology, 10*(1), 51–57. https://doi.org/10.22214/ijraset.2022.39743

Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., ... & Choi, S. W. (2024). Data preprocessing techniques for ai and machine learning readiness: Scoping review of wearable sensor data in cancer care. *JMIR mHealth and uHealth*, *12*(1), e59587.

Pk, A., & V, P. (2013). Review on feature selection techniques of DNA microarray data. *International Journal of Computer Applications, 61*(12), 39–44. https://doi.org/10.5120/9983-4814

Shafique, R., Rustam, F., Choi, G. S., Díez, I. D. L. T., Mahmood, A., Lipari, V., ... & Ashraf, I. (2023). Breast cancer prediction using fine needle aspiration features and up sampling with supervised machine learning. *Cancers*, *15*(3), 681.

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., Cheung, N.-M., Nguyen, H. L. T., Ho, C. S. H., & Ho, R. C. M. (2019). Characterizing artificial intelligence applications in cancer research: A latent Dirichlet allocation analysis. *JMIR Medical Informatics, 7*(4), e14401. https://doi.org/10.2196/14401

Turgut, S., Dagtekin, M., & Ensari, T. (2018). Microarray breast cancer data classification using machine learning methods. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 1–3. https://doi.org/10.1109/EBBT.2018.8391468

Ravishankar, M., & Varalatchoumy, M. (2017, December). Four novel approaches for detection of region of interest in mammograms—A comparative study. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 261-265). IEEE.

Ray, P. P., Dash, D., & De, D. (2017). A systematic review of wearable systems for cancer detection: current state and challenges. *Journal of Medical Systems*, *41*, 1-12.

Yao, Y., Lv, Y., Tong, L., Liang, Y., Xi, S., Ji, B., ... & Yang, J. (2022). ICSDA: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data. *Briefings in bioinformatics*, *23*(6), bbac448.