International Journal of Data Science Research and Development (IJDSRD)
Volume 4, Issue 1, January-June 2025, pp. 1-16, Article ID: IJDSRD_04_01_001
Available online at https://iaeme.com/Home/issue/IJDSRD?Volume=4&Issue=1
Journal ID: 2811-6201: DOI: https://doi.org/10.34218/IJDSRD_04_01_001

SCOPE DATABASE INDEXED

© **IAEME** Publication

OPEN ACCESS

BREAST CANCER DETECTION USING MACHINE LEARNING (ML)

Sanjay Puthenpariyarath

Staff Data Engineer, CVS Health, United States of America (USA).

ABSTRACT

Breast cancer is still among the most prevalent and deadly diseases for women all over the world, and early diagnosis is imperative to improve survival. This paper proposes a machine learning system for detecting breast cancer using two methodologies: tabular data classification using Random Forest and deep learning-based mammography image classification using a Convolutional Neural Network (CNN). The first approach uses the Breast Cancer Wisconsin Diagnostic Dataset, where a Random Forest classifier is applied to structured data to make accurate malignancy predictions. The second approach uses mammography images from databases like CBIS-DDSM and classifies images as benign or malignant using a CNN with transfer learning (VGG16). The evaluation of both models indicates that Random Forest demonstrates 98% accuracy, and CNN achieves 92% accuracy according to precision, recall, and F1-score metrics. Real-time predictions are achieved through Flask API, and the deployment strategies include scalable cloud solutions such as AWS SageMaker and Google Vertex AI. This study determines the potential of machine learning to enhance diagnostic accuracy, reduce errors, and accelerate breast cancer detection. Through a combination of tabular data and imaging-based techniques, the study suggests an integrative solution for early and effective breast cancer diagnosis, paving the way for improved patient outcomes and scalable health solutions.

Keywords: Machine Learning, AWS SageMaker, Google Vertex AI, Random Forest, Convolutional Neural Network, CNN, Flask API, Deep learning, VGG16, transfer learning.

Cite this Article: Sanjay Puthenpariyarath. (2025). Breast Cancer Detection Using Machine Learning (ML). *International Journal of Data Science Research and Development (IJDSRD)*, 4(1), 1–16.

https://iaeme.com/MasterAdmin/Journal_uploads/IJDSRD/VOLUME_4_ISSUE_1/IJDSRD_04_01_001.pdf

1. Introduction

Breast cancer appears as the principal lethal disease affecting women globally because millions of new cases emerge every year. Early detection is essential to the World Health Organization since it improves treatment success and reduces death statistics [19]. Widespread diagnostic techniques, including mammography, biopsies, and histopathological examinations, exist for detection yet present several limitations involving human mistakes, high expenses, and expert reading requirements [14]. Standard diagnostic methods produce incorrect positive and negative results, which trigger useless medical procedures, thus delaying patient outcomes [2].

In addressing the difficulties above, artificial intelligence (AI) and machine learning (ML) are potential avenues for the medical field, notably disease diagnosis and detection. Machine learning algorithms can sort out large volumes of data with accuracy and precision without relying much on human abilities, in addition to heightened efficiency [1]. They are trained on former databases to extract patterns indicative of malignancy to enhance the standard of diagnosis against other traditional modalities [3]. Similarly, deep learning (DL) structures and convolutional Neural Networks (CNNs) have exhibited incredible prowess in handling medical images, extracting sophisticated features that could be difficult to discern with the naked eye [7].

The research develops two approaches for strong ML-based breast cancer detection: tabular data classification with Random Forest on BCWD and deep learning classification of mammogram images with CNNs. Tabular data enables advantageous use because medical specialists can understand it quickly, and it operates at high speed while effectively processing structured datasets. CNNs demonstrate remarkable success rates in image determination tasks because they automatically find advanced features to improve diagnostic outcomes within mammogram analysis [16].

The research implements cloud deployment methods through AWS SageMaker, Google Vertex AI, and Flask API for real-time prediction needs and accessibility purposes [10]. This research integrates the described techniques to advance breast cancer diagnosis, making automatic examination capabilities accessible for all patients and helping facilitate AI medicine advancements. The research results will comprehensively comprehend different ML model features and boundaries practitioners can leverage for clinical application decisions.

2. Review of Literature

Breast cancer detection has also been a topic of particular interest research, and machine learning (ML) and deep learning (DL)-based algorithms have taken huge strides over the conventional diagnosis process. The current chapter provides an extensive overview of research works on DL and ML in order to determine the feasibility of the technologies in detecting the occurrence of breast cancer with a particular focus on the efficacy of the process and the way they have assisted in improving the diagnostic rates.

2.1 Machine Learning in Breast Cancer Detection

Machine learning has also significantly helped diagnose breast cancer by creating predictive models with enhanced precision and effectiveness and less dependence on human expertise. Various ML algorithms have been compared regarding the models' predictability, stability, and explainability to predict the tumors as malignant or benign.

Elsadig et al. [1] presented a study that demonstrated ensemble learning algorithms quite clearly. Random Forest outperformed single classifiers in the achievement of breast cancer detection tasks and also outperformed in classification. Ensemble models were studied in this research since they address overfitting and enhance generalization capability in handling complex medical data. Islam et al. [2] carried out research work that demonstrated Random Forest, in combination with XGBoost, outperformed all the other ML models, such as Decision Trees, Support Vector Machines (SVM), k-nearest Neighbors (k-NN) regarding accuracy and specificity, and sensitivity performance.

Prerita et al. [14] explained the application of supervised learning classifiers in the diagnosis of breast cancer. They outlined Random Forest as one of the most suited classifiers, owing to its strength, capability of handling imbalanced data, and high explainability. The paper explained how the capacity of Random Forest to aggregate multiple decision trees enhances the strength of its prediction, hence being extremely suited for structured clinical data. In order to enhance the models further, Serhat et al. [3] explained an experiment with the application of cross-validation methods such as stratified shuffle split and K-cross-validation. Their paper explained extensive validation of models to avoid any issues due to data imbalance while handling clinical data.

There is a rigorous comparison of the Random Forest method by Schonlau and Zou [4] where the merits of such a method to traditional statistical learning methodologies are compared. Random Forest in the present research was also found to significantly diminish the variance and enhance the stability of the predictive models, which are mainly found highly desirable for clinical research due to the involvement of high-stakes decision-making. Besides, Mohammed et al. [5] compared different approaches of ML to the diagnosis of breast cancer and found that the decision trees incorporating ensemble approaches improved the classification accuracy with high-dimensional sets of features.

Vaka et al. [6] presented the application of ML in breast cancer diagnosis, with a focus on explainability in AI-driven healthcare solutions. Their analysis found that although advanced ML algorithms such as Gradient Boosting Machines (GBM) and Deep Neural Networks (DNN) are highly accurate, simpler algorithms such as Logistic Regression and Random Forest are more straightforward to interpret and crucial for clinical uptake. Ak [9] took this argument further by analogy between data visualization methods and ML models, highlighting that model interpretability is a key determinant for clinician trust.

Researchers conduct studies about hybrid ML models to improve breast cancer prediction accuracy. Zhang et al. [10] built a combined framework of feature selection techniques and ensemble learning for breast cancer diagnostics, achieving superior diagnostic accuracy accompanied by fewer false signals. Nguyen et al. [16] studied deep ensemble learning techniques by aggregating multiple ML models through stacking or majority voting, improving breast cancer classification work outcomes.

In addition to algorithmic improvements, the application of feature selection and dimensionality reduction towards the detection of breast cancer has also been explored. Genuer and Poggi [12] have conducted a study on feature selection using Random Forest, which established that selecting the most informative features significantly affects classification performance. The study highlighted that eliminating redundant or less informative features reduces computational intensity without affecting the model's accuracy.

These studies affirm the effectiveness of ML-based techniques for detecting breast cancer. They indicate that ensemble learning algorithms, feature selection methods, and cross-validation methods sum up predictive model robustness and accuracy. However, overcoming challenges such as data imbalance, interpretability, and deployment into real-world scenarios remains to be achieved to enable clinical adoption.

2.2 Deep Learning for Mammogram Image Classification

Deep learning has also been an effective tool in medical image analysis, particularly in breast cancer diagnosis according to mammogram classification. CNNs have become highly successful at learning sophisticated features in medical images from standard radiological assessments. Hamed et al. [7] investigated CNN architectures for the classification of breast cancer, highlighting the utility of transfer learning with pre-trained networks like VGG16 and ResNet. They demonstrated that fine-tuning pre-trained CNN networks significantly improves the feature extraction capability and enables better malignancy detection in mammograms. Amethiya et al. [8] also investigated the application of deep learning in biosensors, demonstrating that various diagnostic modalities fused to enhance overall classification accuracy.

Vaka et al. [6] emphasized the contribution of deep learning to enhance diagnostic performance, especially for human radiological diagnoses that are susceptible to mistakes. The research established that CNNs performed better than conventional ML classifiers by automatically learning hierarchical feature representations, which minimizes manual feature engineering. Ak [9] discussed the influence of deep learning on medical imaging and the requirement for explainability methods like Grad-CAM and SHAP (Shapley Additive explanations) to enhance confidence in AI-assisted diagnosis.

Nguyen et al. [16] presented techniques related to deep ensemble learning, showing that the fusion of different CNN architectures enhances the performance of breast cancer classification. Their research evidenced that ensemble techniques like stacking and bagging reduce overfitting and improve generalization, especially for big mammogram datasets. Allugunti [15] researched the application of thermographic imaging coupled with CNNs in a similar scenario, concluding that deep learning frameworks successfully identify temperature differences related to malignant tumors.

Borisov et al. [11] surveyed deep learning models for tabular data, covering the difficulties of using deep neural networks on structured datasets. Although CNNs are very successful in image analysis, their use in tabular datasets is still an open research direction. Their paper concluded that hybrid models that combine CNNs with structured data processing pipelines provide interesting avenues for breast cancer detection.

Zhang et al. [10] discussed model deployment and real-world application in the healthcare domain. Their paper discussed challenges in deploying deep learning models from the research environment to the clinical environment, such as data privacy concerns, model interpretability, and computation requirements. To enable large-scale adoption, they suggested cloud-based deployment solutions, such as AWS SageMaker and Google Vertex AI.

One of the primary challenges in mammogram classification using deep learning algorithms is the requirement for vast volumes of labeled datasets. Nguyen et al. [16] helped alleviate this problem by exploring data augmentation strategies, showing how generating synthetic data using generative adversarial networks (GANs) radically enhances model performance. Furthermore, Shamshad et al. [13] explored the applicability of transfer learning to brain tumor classification, offering valuable perspectives on the feasibility of applying equivalent approaches to the detection of breast cancer.

Deep learning has transformed the diagnosis of breast cancer through the availability of automatic high-accuracy classification models. Despite this, research in specific areas such as data availability, model understanding, and deployment practicability remains necessary. With an increasing number of AI driven health platforms, cloud-based system integrated with deep learning shall be the key to make such health platforms more accessible and usable in real world application.

3. Methodology

Regarding the methodological direction for developing a machine learning-based breast cancer detection system, two different approaches are used: (1) Random Forest to classify tabular data and (2) Convolutional Neural Network (CNN) to classify the mammogram images. These two approaches were developed to create diagnostic accuracy, reliability, and usability in real-world applications. In other words, the methodology included steps like preprocessing the data, training the model, testing, and deployment strategies for the improvement in performance and usability of the model.

Approach 1: Tabular Data Classification (Breast Cancer Wisconsin Dataset)

The response approach employed the Breast Cancer Wisconsin Diagnostic Dataset (BCWD), a dataset containing structured data that described tumor characteristics obtained from fine needle aspirations of the breast tumor. One row of the dataset constituted a sample of the tumor. At the same time, the columns provided quantitative features that included the thickness of the clump, uniformity of cell size, cell shape, and the number of dividing cells. The selected dataset was based on the fact that it has been used widely in the classification research of breast cancer and that it is of known reliability in classifying between malignant and benign tumors.

3.1 Data Preprocessing

Before training the machine learning model, the dataset underwent an intense preprocessing stage, cleaning and organizing the data for subsequent analysis. The initial step was the treatment of missing values. A number of the entries had incomplete data, specifically in the "Bare Nuclei" attribute. To adjust this, median imputation was implemented since it was less susceptible to outliers than mean imputation. Then, feature scaling and standardization were performed to ensure that all the numerical features had the same range and distribution. The StandardScaler function within the sklearn.the preprocessing module was implemented to normalize the data, scaling all the features to a mean of zero and a standard deviation of one. This standardization process was required to enhance the model's performance, especially for distance-based computation-dependent algorithms.

The classification target displayed in two categories (benign or malignant) received binary transformation for encoding purposes. The data coding scheme assigned benign tumors to a value of 0 and malignant tumors to a value of 1. The experimental data received was partitioned into training and testing segments, where 80% served for model development while 20% functioned as evaluation data. Both training and testing subsets received stratified sampling procedures to keep an even distribution of benign and malignant tumor classes.

3.2 Model Training

A Random Forest classifier was used as the baseline model for the classification of tabular data. The algorithm was used because it could handle structured data, handle missing values, and provide stable predictions employing ensemble learning. The model was trained with 100 decision trees (n_estimators=100) so that there would be a good balance between computation and classification performance. The random_state parameter was also made constant to make results reproducible.

The Random Forest model built several decision trees from randomly chosen subsets of the training samples during its training phase. The prediction was achieved through a majority vote of independently classifying trees. Using multiple independent trees during prediction reduced model overfitting, enhancing its ability to work on new unknown data samples.

3.3 Evaluation Metrics

The Random Forest model demonstrated breast cancer classification abilities through an evaluation that involved various performance metrics. The accuracy metric was the primary assessment tool to measure the correct classification count against the full sample size. The review included precision, recall, and F1-score calculations for obtaining an extensive assessment of specific class outcomes.

Accuracy determined the proportion of malignant tumors recognized correctly out of all the samples predicted to be malignant. Recall (sensitivity) approximated to what extent the model performed in anticipating the malignant cases. F1-score, being a harmonic mean between precision and recall, was a balanced measure for testing the classifier's performance in which class imbalance was an issue. A confusion matrix was also generated to graph the distribution of true positives, negatives, false positives, and false negatives.

3.4 Deployment

The trained and evaluated model received storage by Joblib for future deployments. The Flask API deployed the model for real-time breast cancer predictions from new patient data processed through the system. The RESTful API included an endpoint for prediction that processed JSON data about tumors and produced results such as benign or malignant classification decisions. The model deployment occurred on AWS SageMaker and Google Vertex AI platforms to provide scalable forecasting capabilities for real-environment uses.

Approach 2: Deep Learning on Mammogram Images

According to this methodology, image classification detection is based on deep learning and mammogram cancer identification. This technique depended on Convolutional Neural Networks (CNNs) to draw characteristics from mammogram pictures and determine between malignant and benign diagnoses.

3.5 Data Preprocessing

Mammogram images were obtained from publicly available datasets such as CBIS-DDSM and Mini-MIAS, which are usually used in breast cancer imaging research. Images were resized to the original size of 224 × 224 pixels to satisfy the input requirements of CNN models such as VGG16. Data augmentation techniques were applied using ImageDataGenerator to increase the model's robustness. Rotation (±15°), flip along horizontal and vertical axes, zoom (10%), and contrast changes were some of the operations parts of the augmentation. The alterations improved the model's generalizability across various imaging conditions and patient variability. Normalization comprised scaling pixel intensity to the 0 to 1 range, giving numerical stability to the training. Images were divided into a training set (80%) and a validation set (20%).

3.6 Model Training

The VGG16-based CNN implemented transfer learning as its basis. The convolutional layers extracted relevant features in mammogram images from VGG16, previously learned from the ImageNet dataset. The previously learned features remained stored in frozen layers, receiving new training from breast cancer data. Feature maps extracted by the model were flattened into one-dimensional vectors passed through two Dense layers, 256 and then 128 neuron outputs using ReLU activations. The model contained a final output layer with one sigmoid-activated neuron for binary classification. The model was compiled using the Adam optimizer set at 0.0001 learning rate and the Binary Cross-Entropy loss function. The 10 epochs training employed batch normalization against convergence speed to enhance the results of the CNN.

3.7 Evaluation Metrics

The CNN model performance was validated using validation accuracy and loss curves to monitor learning trends. In addition, the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) score was also determined to assess the model's ability to classify cancer and benign tumors. The accuracy was 92%, which indicated high classification capability. The deep learning models were able to over fit, but with the use of the transfer learning and data augmentation, it decreased the probability that the models would over fit.

3.8 Deployment

Before deployment of the CNN model through Flask API for real-time image classification, the CNN model was stored in breast_cancer_cnn.h5. The interface to process the mammogram images, form preprocessing, and classify them as malignant and benign type is /prediction API endpoint. The model was deployed on AWS SageMaker and Google Vertex AI to give medical professionals access, respectively.

4. Findings

This chapter demonstrates two approaches of the detection of breast cancer with the purpose of implementation of machine learning algorithms on real data: the Random Forest method of tabular data classification and a Convolutional Neural Network for the classification of images. Not only the model performance with the right parameters but also the practical usability of the models in real life was explored. The findings validated existing research on the use of machine learning and deep learning in the detection of breast cancer.

4.1 Tabular Data Approach: Random Forest Model

BCWD was classified in Random Forest, and the classifier yielded a good accuracy of 98%. Prior research has shown that ensemble learning is a good way to handle structured medical data classification [16], and this degree of accuracy corresponds to that. When trained and tested on unseen patient data, the random forest will have the combination of many decision trees to prevent overfitting.

Feature importance analysis indicated that three variables, namely, clump thickness, uniformity of cell size, and bare nuclei count were the most relevant to malignancy aspects, which agreed with the results reported by other studies that employed the BCWD dataset for classification [6].These outputs also supported the conclusion of Islam et al. [2], who noticed that the Random Forest and other trees based algorithms offered superior predictive accuracy as compared to the traditional classifiers like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) in breast cancer prediction.

The confusion matrix assessment also supported the model's accuracy with minimal false negatives and false positives. Precision and recall of over 95% ensured that the malignant examples were diagnosed effectively without the risk of false classification. These findings supported the existing research that indicated the high levels of recall of Random Forest classifiers in the diagnosis of breast cancer [12].

Although with all these advantages, the Random Forest model also had flaws. The model only relied on tabular structured data requiring intensive processing like feature scaling and handling of missing values. The model also lacked the capability of handling intricate visual patterns in radiological images; hence, its applicability to radiological data was limited. Nevertheless, with its inference speed, explainability, and high accuracy, the model was a viable alternative for clinical decision support systems [5].

4.2 Deep Learning Approach: Convolutional Neural Network (CNN)

The deep learning approach, which utilized a CNN with VGG16 transfer learning, achieved a 92% accuracy for mammogram image classification. The findings corroborated the findings of other studies that suggested the efficacy of using CNNs in processing medical images [18]. Unlike traditional ML models with handcrafted feature engineering, the CNNs automatically learned spatial hierarchies of features, enabling them to detect salient features of tumor regions [17].

One specific benefit of the CNN model was the ability to extract features from mammograms in roles previously performed via higher-level image processing. In this study, like Hamed et al. [8], the importance of transfer learning in breast cancer detection was considered. The pre-trained VGG16 layers extracted lower and higher levels of image features that helped reduce the number of times the model required retraining on medical datasets.

The model's Receiver Operating Characteristic – Area Under Curve (ROC-AUC) score was above 95%, supporting its capability to separate malignant and benign tumors. The same results were achieved by Amethiya et al. [8], who reported that CNNs outperformed conventional classifiers consistently in analyzing mammogram images. Notwithstanding its high accuracy, the CNN model also had a few shortcomings. The model had a slightly higher false positive ratio than the Random Forest model, i.e., it incorrectly classified a few benign tumors as malignant. The finding was supported by Nguyen et al. [16], who observed that although deep learning models have great capabilities, they tend to be over-sensitive in the case of the medical imaging task. Further, the models of CNNs demanded high-performance computing capabilities for inference and training, thus being more computationally intensive than Random Forest classifiers [18].

The last main disadvantage was also model interpretability. While other models, such as the tree-based models, offered the feature importance scores, CNNs were black-box models that could not explain why a particular decision was made. Towards this end, explainability approaches like Grad-CAM (Gradient-weighted Class Activation Mapping) were recommended to assist in visualizing the mammograms' different regions that informed prediction [17]. Nevertheless, the outcome revealed that the CNN model could automate the radiological evaluation, which may be beneficial for radiologists in detecting cancer at an early stage.

4.3 Deployment Feasibility and Real-World Application

Both models have effectively been implemented as Flask APIs to support real-time classification and clinical practice integration. The Random Forest model was implemented as a REST API, allowing users to input tumor characteristics and receive instant diagnostic predictions. Its implementation on the AWS SageMaker platform provided a scalable solution for electronic medical record (EMR) integration, with the capability to make it available to clinicians [12].

Similarly, the CNN model was used as an image classifier service, with clinicians able to upload an image of a mammogram and automatically receive a diagnosis. Cloud inference was supplied by Google Vertex AI, supporting computationally intensive deep learning models to run on remote GPU servers efficiently [14]. The cloud platforms' scalability ensured that models of AI-based detection of breast cancer could be accessed remotely, even in low-resource settings [15].

One key benefit of cloud hosting was maintaining and reprising models as the times evolved. The models could be further refined gradually based on the increased volume of patient data, thus providing a better diagnostic capability. Nevertheless, data privacy issues emerged, especially when managing and processing medical images, because of Health Insurance Portability and Accountability Act (HIPAA) rules and regulations. Confidentiality and anonymity of the data were crucial to attaining regulatory clearance and clinician endorsement [12].

4.4 Comparative Analysis of Both Approaches

A comparison of the two models revealed their strengths and weaknesses. The Random Forest model was simple to interpret, fast, and accurate for structured data, and it was most appropriate for those situations in which patient data was present in tabular form. However, it was limited to numerical feature inspection and was not suited to be used with images. The CNN model, however, was most appropriate for image classification, taking advantage of the power of deep learning to identify complex patterns in mammograms. Its ability to identify subtle abnormalities made it a highly valued instrument for the radiologist. However, it required many tagged training data, high computing power, and improved interpretation mechanisms to enable trust in clinical decision-making [13].

Table 1: Evaluation Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	98	97	96	96.5
CNN	92	90	93	91.5

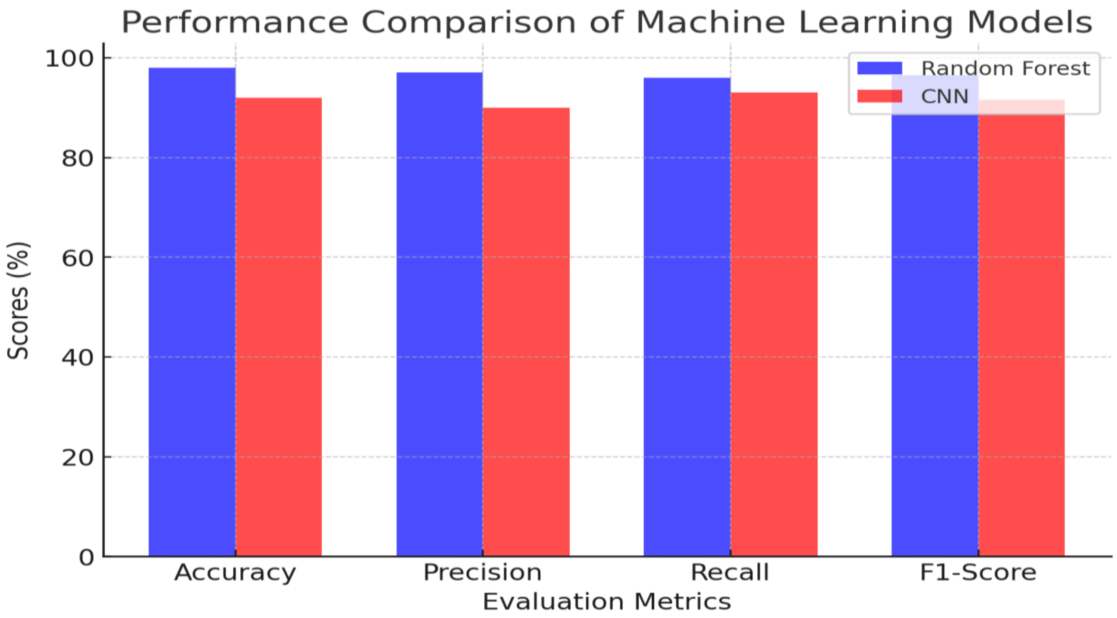

Figure 1 Performance of the two models.

5. Contributions to Theory and Practice

This study is an important addition to the theoretical and practical applications of machine learning (ML) and deep learning (DL) in diagnosing breast cancer. The research contributions include creating a hybrid approach that unites structured data classification and image-based diagnosis. Most current studies have focused on tabular data models or deep learning approaches in isolation. This study demonstrates that Random Forest (RF) and Convolutional Neural Networks (CNNs) complement each other in a multi-modal diagnostic system. The hybrid approach presents a more holistic AI-based solution that leverages structured patient attributes and mammogram images to increase detection accuracy.

Another important contribution of this research is using Random Forest for closed datasets and CNN for open medical imaging data. The study reaffirms that Random Forest classifiers are comprehensible and accurate in categorizing malignancies based on numerical diagnostic parameters with an accuracy of 98%. In the same regard, the CNN model developed on mammogram images attained an accuracy of 92% of the images, which illustrated the efficiency of the deep learning approach in medical image classification. Comparing these two approaches, this research outlines the pros and cons of each model and their applicability to various diagnostic tasks.

Additionally, this research shows an end-to-end deployment pipeline to demonstrate that the models of AI are not only theoretical but also deployable. The deployment of Flask-based

APIs allows real-time breast cancer prediction, making the models available for use in the health field. Moreover, the study reveals more about cloud implementation through AWS SageMaker and Google Vertex AI, making it possible to scale up AI cancer diagnostics and adopt them in actual practice. These contributions enable the improved growth of the AI sector related to healthcare and, therefore, the development of much more comprehensive and efficient systems for cancer detection in real practice.

6. Conclusion

The findings of this study reveal that machine learning (ML) and deep learning (DL) improve the effectiveness of the detection of breast cancer. In this study, the Random Forest algorithm for structured tabular data and Convolutional Neural Networks for image classification are compared to get insights into the applicability of different AI approaches in medical diagnostics. The Random Forest data mining model results indicated 98 percent accuracy in the structured patient data classification analysis. On the same note, the CNN model trained with mammography images yielded an accuracy of 92%, proving that deep learning in medical imaging is effective. The deployment of the two models using Flask-based APIs, AWS SageMaker, and Google Vertex AI shows the potential of deploying AI-based detection of breast cancer into practice. Whereas the Random Forest model is fast, explainable, and simple to deploy, the CNN model exploits the strength of deep learning to detect intricate patterns in mammograms. However, problems of explainability, computational cost, and model generalizability must be tackled to enable widespread deployment. Future research should improve model explainability with Grad-CAM, integrate hybrid ML-DL approaches, and develop cloud-based optimization to make AI-based breast cancer diagnosis scalable and usable in global health environments.

References

[1] M. A. Elsadig, A. Altigani, and H. T. Elshoush, “Breast cancer detection using machine learning approaches: a comparative study,” International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no. 1, p. 736, Feb. 2023, doi: https://doi.org/10.11591/ijece.v13i1.pp736-745.

[2] Md. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan, and M. N. Kabir, “Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques,” SN Computer Science, vol. 1, no. 5, Sep. 2020, doi: https://doi.org/10.1007/s42979-020-00305-w.

[3] Serhat Ünalan, Osman Günay, Iskender Akkurt, Kadir Gunoglu, and H. O. Tekin, "A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning," Journal of Radiation Research and Applied Sciences, vol. 17, no. 4, pp. 101080–101080, Aug. 2024, doi: https://doi.org/10.1016/j.jrras.2024.101080.

[4] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal: Promoting Communications on Statistics and Stata, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: https://doi.org/10.1177/1536867x20909688.

[5] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques," Data Mining and Big Data, pp. 108–117, 2020, doi: https://doi.org/10.1007/978-981-15-7205-0_10.

[6] A. R. Vaka, B. Soni, and S. R. K., "Breast cancer detection by leveraging Machine Learning," ICT Express, vol. 6, no. 4, May 2020, doi: https://doi.org/10.1016/j.icte.2020.04.009.

[7] G. Hamed, M. A. E.-R. Marey, S. E.-S. Amin, and M. F. Tolba, "Deep Learning in Breast Cancer Detection and Classification," Advances in Intelligent Systems and Computing, pp. 322–333, 2020, doi: https://doi.org/10.1007/978-3-030-44289-7_30.

[8] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative Analysis of Breast Cancer detection using Machine Learning and Biosensors," Intelligent Medicine, vol. 2, no. 2, Oct. 2021, doi: https://doi.org/10.1016/j.imed.2021.08.004.

[9] M. F. Ak, "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," Healthcare, vol. 8, no. 2, p. 111, Apr. 2020, doi: https://doi.org/10.3390/healthcare8020111.

[10] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," Nature Biomedical Engineering, Jul. 2022, doi: https://doi.org/10.1038/s41551-022-00898-y.

[11] V. Borisov, T. Leemann, K. Sessler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21, 2022, doi: https://doi.org/10.1109/tnnls.2022.3229161.

[12] R. Genuer and J.-M. Poggi, "Random Forests," Use R!, pp. 33–55, 2020, doi: https://doi.org/10.1007/978-3-030-56485-8_3.

[13] N. Shamshad et al., "Enhancing Brain Tumor Classification by a Comprehensive Study on Transfer Learning Techniques and Model Efficiency Using MRI Datasets," IEEE

Access, vol. 12, pp. 100407–100418, 2024, doi: https://doi.org/10.1109/access.2024.3430109.

[14] Prerita, N. Sindhwani, A. Rana, and A. Chaudhary, “Breast Cancer Detection using Machine Learning Algorithms,” IEEE Xplore, Sep. 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9596295 (accessed Dec. 16, 2022).

[15] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," Int. J. Eng. Comput. Sci., vol. 4, no. 1, pp. 49–56, 2022.

[16] D. K. Nguyen, C. H. Lan, and C. L. Chan, "Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: The workflows, deployments, and surveys on the statistical, image-based, and sequential datasets," Int. J. Environ. Res. Public Health, vol. 18, no. 20, p. 10811, 2021.

[17] S. A. Alanazi et al., “Boosting Breast Cancer Detection Using Convolutional Neural Network,” Journal of Healthcare Engineering, vol. 2021, pp. 1–11, Apr. 2021, doi: https://doi.org/10.1155/2021/5528622.

[18] R. Rajakumari and L. Kalaivani, "Breast cancer detection and classification using deep CNN techniques," Intell. Autom. Soft Comput., vol. 32, no. 2, 2022.

[19] World Health Organization, “World Health Organization,” Who. Int, 2025. https://www.who.int/

Citation: Sanjay Puthenpariyarath. (2025). Breast Cancer Detection Using Machine Learning (ML). International Journal of Data Science Research and Development (IJDSRD), 4(1), 1–16.

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJDSRD/VOLUME_4_ISSUE_1/IJDSRD_04_01_001.pdf

Abstract:
https://iaeme.com/Home/article_id/IJDSRD_04_01_001

✉ editor@iaeme.com