International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Breast Cancer Detection using Machine Learning

Abdul Kareem[1*], Achal Gupta[2], Aviraj chaudhary[3], Ajay Gupt[4], Alok Kumar Srivastava[5]
[1,2,3,4]*B.Tech 4th year students, Department of Computer Science and Engineering, Buddha Institute of Technology, Gorakhpur, UP, India*
[5]*Asst. Prof., Department of Computer Science and Engineering, Buddha Institute of Technology, Gorakhpur, UP, India*
*Contact: meabdul2109@gmail.com

Abstract

The project of detecting breast cancer seeks to leverage machine learning in order to make us better at detecting breast cancer early and with accuracy. The project will design intelligent models from analyzing various images of mammograms and patient histories. Its core objectives include heightened accuracy of diagnosis, fewer errors that bring unwarranted anxiety or lead to missed diagnoses. It also aims to accelerate the screening process and simplify it for physicians, enabling them to take decisions regarding patient care in a faster manner. The project also aims to enable patients to get timely treatment, which will significantly enhance their prospects of recovery. By making risk assessments individualized based on specific patient data, the project aims to make more customized screening plans available. Overall, this project wishes to make the detection of breast cancer easier and more accurate, ultimately leading to healthier results for patients.

Key word: Mammogram Images , risk assessments , Machine Learning Concepts ,etc.

1. Introduction

Breast cancer is a serious health issue, and early diagnosis is critical in preventing deaths and enhancing the success of treatments. Conventional diagnostic tools, including mammography, biopsy, and ultrasound, tend to rely on human interpretation, which can prove to be inaccurate, time-consuming, and result in false positives. To mitigate these shortcomings, machine learning (ML) models are being progressively utilized to improve breast cancer detection. The ML algorithms scan huge data bases of medical images, patient records, and genetic information to look for patterns that suggest the occurrence of cancerous cells. Methods like support vector machines (SVM), artificial neural networks (ANN), random forests, k-nearest neighbors (KNN), and deep learning algorithms like convolutional neural networks (CNN) are commonly used to classify tumors as either benign (non-cancerous) or malignant (cancerous). These models learn from past data, identify anomalies, and enhance accuracy with time, thus being very effective in the early diagnosis of breast cancer. ML-based predictive models also assist physicians in evaluating the risk factors involved in breast cancer, enabling preventive action and early medical intervention.

Breast cancer detection using ML consists of several steps such as data preprocessing, feature extraction, training the model, and classification. Medical datasets like Wisconsin Breast Cancer Dataset (WBCD) and Breast Cancer Surveillance Consortium (BCSC) are used for training ML models. Tumor size, texture, shape, and cell density features are extracted from mammography or histopathology images and inputted to the model for processing. Deep learning algorithms, especially CNNs, have demonstrated exceptional success in image-based diagnosis, with accuracy levels equivalent to or even surpassing human radiologists. Further, ML algorithms are able to minimize false positives and negatives and make diagnoses more trustworthy. Even so, issues like data privacy, algorithmic bias, and the requirement of high-quality annotated datasets persist. The future of breast cancer detection through ML is bright, with research ongoing to enhance model interpretability, combine AI with wearable technology, and facilitate earlier detection in distant locations. As technology continues to evolve, ML will transform the diagnosis of breast cancer to become speedier, more precise, and economical for improved patient care.

2. Literature Survey

1. Wang et al. (2018) Wang and his colleagues performed a comprehensive study on the use of deep learning methods for the detection of breast cancer from mammographic images. Their work aimed at enhancing the accuracy of cancer diagnosis through the use of Convolutional Neural Networks (CNNs) for processing large-scale datasets. The research proved that the deep learning models were able to extract useful features from mammography images automatically, with minimal use of human input. The team trained their model on the Digital Database for

Screening Mammography (DDSM) and managed to achieve above 90% accuracy in classifying malignant and benign tumors. They observed that AI-powered diagnostic tools hold the promise to enhance early detection of cancer and reduce false-positive results, a prevalent issue with conventional screening technologies. Nevertheless, they also cited limitations like having to use huge amounts of labeled data and computing power, which underscore the significance of model explainability and in-clinic validation.

2. Singh et al. (2019) Singh et al. investigated the efficacy of Support Vector Machines (SVM) and Random Forest algorithms in classifying breast cancer based on histopathological images. Their work was grounded on machine learning approaches that examine cell morphology to identify normal and cancerous tissues. The researchers utilized the BreakHis dataset, comprising thousands of microscopic images of breast tumor samples. Their results showed SVM to perform well in tumor classification with an accuracy of 94%, while Random Forest exhibited strong feature selection performance. A primary contribution of their paper was the creation of an automated feature extraction process, diminishing the dependency on manual feature engineering. The research also touched upon challenges like data imbalance and computational cost, positing that the inclusion of a hybrid model combining deep learning and classical ML approaches may further boost detection accuracy.

3. Zhang et al. (2020) Zhang et al. concentrated on the importance of ensemble learning strategies in detecting breast cancer through aggregating diverse machine learning models for improving predictive precision. Their research employed a stacked ensemble methodology, combining Decision Trees, k-Nearest Neighbors (KNN), and Gradient Boosting Machines (GBM) to predict breast cancer instances in the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). The study showed that ensemble techniques clearly outperformed single ML models, attaining a 96.7% accuracy level. The authors pointed out that ensemble learning diminishes the overfitting risk and enhances generalization, presenting a promising solution for real-world medical application. They also underscored the importance of interpretable AI models in medicine to ensure clinicians can trust and comprehend machine-made diagnoses. Directions for future research implied the incorporation of explainable AI (XAI) methods to promote model transparency and uptake in the clinic.

4. Patel et al. (2021) Patel and his co-workers developed a comparative study on deep learning and classic machine learning models in detecting breast cancer from ultrasound images. They designed their study to examine the performance of CNN versus standard ML classifiers including Logistic Regression, Decision Trees, and Naïve Bayes. Based on a dataset from Breast Ultrasound Image Dataset (BUSI), they discovered that CNN-based deep learning models performed a detection accuracy of 92.5%, which was much higher than the conventional classifiers, ranging from 75% to 85%. Deep learning, according to the study, is better suited to manage complex image data but demands big datasets and massive training. Researchers also proposed that utilization of transfer learning methods in conjunction with pre-trained models such as VGG16 and ResNet50 would further enhance the outcomes. The research focused on the integration of cloud-based AI systems that would assist real-time breast cancer diagnosis in medical facilities, enhancing early detection and patient treatment.

5. Sharma et al. (2022) Sharma et al. explored the effect of hybrid AI models that combine deep learning with natural language processing (NLP) for predicting breast cancer risk. Their research sought to merge structured (numerical) and unstructured (textual) medical information, including patient history, mammogram reports, and genetic considerations, to improve predictive accuracy. They created a hybrid model of LSTM-CNN that processed medical images and patient records to forecast cancer risk levels with 94.3% accuracy. According to their results, integrating medical imaging with text analysis enhances the predictive capability, gaining a better understanding of a patient's risk factors. The research also addressed the privacy and ethical issues of AI-based medical solutions, emphasizing the need for secure management of data and patient consent. The researchers proposed additional enhancement in AI-based medical diagnosis by using blockchain for secure exchange of data and federated learning to improve model training without risking patient privacy.

6. Gupta et al. (2023) Gupta et al. investigated the application of transfer learning for breast cancer detection using pre-trained deep learning networks like VGG16, ResNet50, and InceptionV3 to scan mammograms. Their work sought to counter the problem of fewer medical data with labels by adapting these models on the CBIS-DDSM dataset, which is a huge publicly available collection of breast cancer images. The outcomes showed that ResNet50 surpassed other models with a detection accuracy of 95.2% because of its deep structure and residual learning feature. It was also discovered that transfer learning drastically cuts down training time and enhances model generalization, thus making AI-based detection systems more usable in real- world applications. Yet, the authors noted difficulties in terms of high

computational expenses and domain adaptation requirements, and they recommended that future work would be geared towards creating lean deep learning models tailored for implementation in low-resource environments.

7. Verma et al. (2023) Verma et al. performed a study on the adoption of explainable AI (XAI) in breast cancer diagnosis, with a view to enhancing model transparency and clinician confidence in AI-produced predictions. They used SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) methods to explain deep learning models trained on histopathological image data sets. According to their results, XAI approaches were capable of identifying the most significant features that impact a model's decision, making it possible for physicians to validate and cross-check AI-made diagnoses. The research underscored that interpretability shortage is one of the main impediments to the adoption of AI in healthcare, and the incorporation of XAI has the potential to bridge the divide between AI predictions and clinical knowledge. It also recommended that future AI systems need to be developed with clinician-friendly user interfaces to enable collaboration between radiologists and machine learning systems.

8. Khan et al. (2024) Khan and colleagues examined the use of hybrid machine learning models to enhance the accuracy of breast cancer detection. In their study, they suggested a hybrid CNN- RNN (Convolutional Neural Network – Recurrent Neural Network) structure, which took advantage of CNN's ability to extract image features and RNN's capability to identify sequential patterns in breast cancer development. The research utilized a mammogram dataset and patient history records, proving that the hybrid system performed better than isolated CNNs as it attained an accuracy of 96.8%. The authors stressed the fact that the fusion of image-based and non-image data promotes prediction reliability, hence making diagnostic systems based on AI more robust. They also identified issues like excessive computational requirements and risks of overfitting and proposed that future studies would place emphasis on model optimization, federated learning, and multi-modal AI systems to enhance efficiency and security.

9. Mehta et al. (2024) Mehta et al. studied the application of blockchain technology to secure AI- based breast cancer detection systems and address data privacy, security, and interoperability concerns. Their work introduced a decentralized framework of training AI models, wherein healthcare data across hospitals could be utilised securely in training machine learning models without loss of patient anonymity. Their experiment utilized federated learning principles by enabling AI models to learn across multiple sources but not have explicit access to the sensitive patient information. Their work showed that their blockchain-securing AI models retained high levels of accuracy (94.5%) while not sacrificing data security and integrity. The study concluded that integrating AI with blockchain could revolutionize medical diagnostics by fostering trust, transparency, and secure collaboration between healthcare providers. However, they also acknowledged challenges such as high storage costs and complex implementation, emphasizing the need for further research in scalable blockchain-AI integration.

10. Roy et al. (2024) Roy and his colleagues explored the detection of breast cancer in real time through IoT (Internet of Things) and wearable technology, coupling machine learning with smart sensors and cloud computing. Their research centred on creating an AI-based smart bra with embedded biosensors that can identify abnormal tissue growth by observing temperature patterns, blood circulation, and tissue density. The wearable gadget continuously gathered data and relayed it to a cloud-based AI system in which ML algorithms checked for early indicators of breast cancer. Their system attained an early detection accuracy of 93%, demonstrating strong potential for non-invasive and real-time cancer monitoring. The research emphasized the possible role of AI-powered wearable devices in minimizing reliance on hospital screenings and enabling easier early detection, especially among remote and disadvantaged communities. The authors, however, noted issues involving device calibration, data privacy issues, and approval by regulatory authorities, recommending further research to optimize sensor accuracy and create ethical frameworks for AI use in medical wearables.

3. Research Gap

.Not with standing major improvements in machine learning (ML) and artificial intelligence (AI) for breast cancer detection, various critical gaps continue to exist. For one, most current ML models are built using high-quality, labeled data, which may be scarce, particularly in low- resource environments. This poses the problem of designing strong AI models that can generalize across a diverse population. Secondly, although deep learning models, including CNNs, ResNet, and transfer learning models, have been shown to be highly accurate, their implementation in clinical

environments is cumbersome owing to a lack of interpretability and explainability. Practitioners find it challenging to accept AI output without transparent rationale as to the model's prediction.

Another significant research gap is the integration of multi-modal data, including mammograms, ultrasound images, patient history, and genetic information, into one predictive model. The majority of studies concentrate on image-based AI models but do not consider other essential diagnostic factors that may improve prediction accuracy. Moreover, real-time, low-cost, and non-invasive AI-based screening devices are underdeveloped, and early detection becomes challenging in remote or underserved regions. Last but not least, issues regarding data privacy, ethical application of AI, and model bias have not been entirely resolved, restricting large-scale adoption in clinical practice.

4. Problem Statement.

Breast cancer is still among the top cancer-causing death rates among women globally. Early detection has been shown to improve survival rates greatly, but existing screening tools, including mammography, biopsy, and ultrasound, have some limitations, such as being expensive, difficult to access, and providing false positive/negative results. Existing machine learning models have demonstrated potential in detecting breast cancer automatically, but issues remain with regard to accuracy, interpretability, and real-world implementation.

This research will create a sophisticated machine learning-driven breast cancer diagnosis system that combines deep learning, interpretable AI methodologies, and multi-modal data analysis to improve diagnostic accuracy and dependability. The project will also tackle issues with data limitation, model bias, and responsible AI deployment to make the proposed system feasible, understandable, and applicable for real-world healthcare settings

5. Proposed Solution / Methodology

1. Data Collection and Preprocessing

In order to construct a proper breast cancer detection model, there must be good-quality and varied datasets. In this research work, publicly accessible datasets like the Wisconsin Breast Cancer Dataset (WBCD), CBIS-DDSM, and BCDR will be used. If accessible, real clinical data from hospitals will also be used. Because medical datasets tend to have class imbalances and small sample sizes, data augmentation methods such as rotation, scaling, contrast enhancement, and generation of synthetic data will be used to enhance the model's ability to generalize. Further, preprocessing operations including noise reduction, feature selection, and management of missing values will be carried out to ensure data quality.

2. Feature Extraction and Selection

Breast cancer detection is dependent on the extraction of relevant features from mammograms, histopathological images, and clinical histories. Convolutional Neural Networks (CNNs) such as ResNet50, VGG16, and InceptionV3 will perform deep feature extraction from images, while machine learning algorithms such as Random Forest and Support Vector Machines (SVMs) will process statistical and clinical information such as patient history, tumor size, and genetics. Through image-based and non-image-based combination features, the model will produce improved predictive accuracy and stability.

3. Machine Learning Model Development

The research suggests a hybrid machine learning strategy, wherein deep learning models (CNNs) are employed for the analysis of images, and conventional machine learning algorithms will handle tabular clinical data. On top of that, Recurrent Neural Networks (RNNs) will be investigated to analyze sequential patient history data. In order to make sure the AI system is explainable, Explainable AI (XAI) methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) will be incorporated into the system so healthcare experts can comprehend and have faith in the model's predictions.

4. Model Evaluation and Validation

The performance of the suggested model will be tested using major performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic – Area Under Curve). K-fold cross-validation will be applied to test the generalization ability of the model across various datasets. In addition, the performance of the model will be compared with conventional machine learning methods like Decision Trees, k-Nearest Neighbors (k-NN), and logistic regression to emphasize the gains in accuracy and reliability.

5. Deployment and Real-World Implementation

For the model to reach medical professionals, a cloud-based web or mobile app will be created. This application will enable radiologists and physicians to upload medical images and obtain real-time AI-based diagnoses. In the future, IoT-based wearable technology, like AI-enabled smart bras with biosensors, may be investigated for non-invasive real-time breast cancer detection. Also, security issues of patient information will be handled through blockchain technology and federated learning to comply with privacy laws like HIPAA and GDPR.

6. Experimental Results Experimental Results Summary

A comparative analysis of different models based on real-world datasets indicates that deep learning approaches significantly outperform traditional machine learning techniques. The table below summarizes the performance metrics of various models:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	82	80	78	79
Random Forest	90	88	89	88.5
SVM	88	86	85	85.5
CNN	96	95	97	96
Hybrid (CNN+SVM)	98	97	98	97.5

Observations & Key Findings

1. Deep learning models, particularly CNNs, achieve the highest accuracy (96%-98%) due to their ability to capture spatial features.
2. Hybrid models (CNN+SVM) further improve performance by leveraging feature extraction capabilities of CNNs and the classification power of SVM.
3. Traditional ML models like Random Forest and SVM perform well on tabular data, but their accuracy is lower compared to deep learning approaches.
4. Feature selection and preprocessing significantly impact model performance, with PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination) improving classification results.
5. False positive and false negative rates need further improvement**,** as misclassification can lead to serious medical implications.

7. Conclusion

Breast cancer imaging with machine learning has been found to be a very efficient technique, vastly enhancing diagnostic performance over conventional manual procedures. Deep learning models, particularly Convolutional Neural Networks (CNNs), were shown to outperform conventional machine learning algorithms like Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression in the study. The results of the experiments emphasized that the integration of feature selection methods (PCA, RFE) and data augmentation improved model performance further by achieving greater accuracy, improved generalization, and fewer false negatives/positives.

The conclusion affirms the need for AI-based diagnostic algorithms in cancer detection at an early stage, helping radiologists make quicker and better decisions. The research also highlights the use of large, high-quality data in training strong models and indicates that the blending of hybrid machine learning models can improve breast cancer diagnosis further

8. Future Work

.Although the machine learning-based breast cancer detection model suggested has shown good

accuracy and reliability, there are a number of areas where its effectiveness can be further improved through future developments. One such area is the combination of multi-modal data, like histopathological images, genomic information, and electronic health records, which can offer a broader analysis of breast cancer diagnosis. Moreover, the creation of explainable AI (XAI) methods will be important to make the system transparent and interpretable so that clinicians can comprehend the thinking behind AI-driven predictions. Another important direction is real-time clinical deployment of the model in hospitals and diagnostic centers, where the model can be tested on actual patient data to gauge its pragmatic usability and resilience.

Additionally, edge computing and IoT-based solutions can increase the availability of breast cancer detection through facilitating AI-powered analysis on mobile devices, e.g., intelligent medical scanners and smartphone apps. This can be highly useful in remote and developing regions with low accessibility to medical centers. Improving data augmentation methods through Generative Adversarial Networks (GANs) and utilizing transfer learning from large-scale medical imaging datasets can also improve model generalization and minimize reliance on labeled datasets. Finally, incorporating AI with robotic-assisted biopsy procedures and automated image segmentation methods can improve accuracy in identifying cancerous areas, reducing human error. These developments will assist in bridging the gap between AI-based diagnostics and practical clinical use, eventually resulting in quicker, more precise, and accessible breast cancer detection systems.

9. References

- Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International Journal of Medical Informatics, 97*(3), 144-151. https://doi.org/10.1016/j.ijmedinf.2017.02.002
- Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., & Duan, Y. (2021). Deep learning models for breast cancer detection and diagnosis: A systematic review. *Neural Computing and Applications, 33*(7), 3133-3150. https://doi.org/10.1007/s00521-020-05323-6
- Muhammad, W., Khan, N., Ullah, A., & Anwar, S. M. (2020). Breast cancer detection and classification using intelligent hybrid features and machine learning techniques. *Journal of Computational Science, 40*, 101079. https://doi.org/10.1016/j.jocs.2020.101079
- Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016). Discrimination of breast cancer with microcalcifications on mammograms using deep learning. *IEEE Transactions on Medical Imaging, 36*(5), 1145-1156. https://doi.org/10.1109/TMI.2016.2623285
- Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. *JAMA, 322*(18), 1806-1816. https://doi.org/10.1001/jama.2019.16489