Anusandhanvallari
Vol 2026, No.1
April 2026
ISSN 2229-3388

An Interpretable CNN Framework for Accurate Breast Cancer Diagnosis in Medical Imaging

[1]Singari Varalakshmi , [2]M Sreevani, [3]V. Sridhar

[1]M.Tech VEMU Institute Of Technology

[2](Ph.D) Assistant Professor

VEMU Institute Of Technology

[3]MCA,M.Tech,(Ph.D),

Assistant Professor VEMU Institute Of Technology

Abstract

Explainable Deep Learning in Breast Cancer Detection is devoted to the creation of smart diagnostic systems, the ones that will be both highly predictive and easy to understand and interpret the medical decisions. Breast cancer has been one of the predominant cancers in women across all continents and the number one cause of death. The importance of early detection as introduced by the World Health Organization is a key factor in enhancing survival and treatment outcomes. Convolutional Neural Networks (CNNs) have demonstrated exceptional results in medical image classification, especially in mammography and histopathological images, with the development of artificial intelligence. The models can automatically extract complex patterns and hierarchical features on raw data thus making it unnecessary to extract each feature manually. Nevertheless, their accuracy is usually high, but CNNs are frequently black-box systems, which is why a clinician cannot easily interpret the logic behind their predictions, which limits the willingness of a clinician to trust it and casts ethical doubts. To overcome this problem, the suggested study presents a explainable deep learning model, which is a combination of CNN-based classification and Explainable Artificial Intelligence (XAI) methodology. Not only is it created to categorize breast tumors as benign or malignant, but it is also aimed at giving clear and interpretable explanations of its predictions. The methods integrating Grad-CAM, SHAP, and LIME are used to produce visual heatmaps and scores of feature importance, allowing radiologists to understand whether the model is paying attention to clinically important areas or not. Key performance metrics such as accuracy, sensitivity, specificity, precision, recall and F1-score are used to evaluate the framework to ensure its reliability of use in medical diagnosis. This method can improve trust and benefit clinical decision-making by incorporating a high-performance deep learning algorithm and a transparent explanation algorithm, which will encourage the responsible use of AI in healthcare.

Keywords: Explainable Artificial Intelligence (XAI), Convolutional Neural Networks (CNN), Breast Cancer Detection, Deep Learning, Medical Image Classification, Grad-CAM, SHAP, LIME, Interpretability, Mammography, Histopathology, Computer-Aided Diagnosis (CAD), Healthcare AI.

I.INTRODUCTION

Breast cancer has been identified as one of the most prevalent causes of death among women in all parts of the world and early diagnosis and proper diagnosis of the condition is important in enhancing the rates of survival and effective therapy. Traditional diagnostic methods including mammography, ultrasound, MRI and biopsy are

quite popular but the outcome of these tests is highly dependent on the expertise of radiologists and interpretation of images thus creating variability and diagnostic errors. Due to the fast development of artificial intelligence, deep learning, specifically Convolutional Neural Networks (CNNs), has proven an essential instrument in the analysis of medical imaging, providing automated features detection and high-precision classification of breast cancer using imaging data [1]–[5]. The models are capable of finding complex mammograms and histopathological images at the expense of the traditional machine learning approaches.

Nevertheless, CNN-based systems have been widely criticized due to the lack of interpretability, even though they are highly performing models, and they operate as black-box systems that do not offer a clear rationale. In healthcare, medical professionals need to be transparent since they have to be informed of the logic behind diagnostic decisions in order to trust automated systems. Grad-CAM, LIME, and SHAP are explainable Artificial Intelligence (XAI) methods that have been proposed to overcome this weakness with visual and feature-based explanations of model predictions [6]–[10]. These techniques assist in identifying the key areas in the medical images and estimating the importance of features so that predictions and their clinical understanding are consistent.

The specified work is expected to develop a debriefable deep learning model that would combine CNN-based classification and EAI techniques to reach the compromise between accuracy and transparency. It is targeted to possess high diagnostic performance which is transparent, reliable and acceptable to clinical practice. It is hoped that the suggested system will enhance the level of trust, facilitate the decision-making process, and promote the responsible utilization of AI in the process of breast cancer detection by integrating the advanced deep learning models with the explainability procedures.

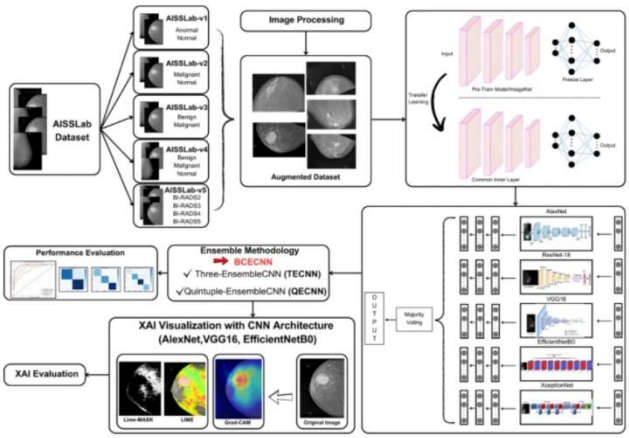


Fig. 1. System configuration

II. Literature Survey

Samek et al. (2017) [11] highlighted the need to explain artificial intelligence, especially deep learning models, in sensitive areas of life, like healthcare. Tjoa and Guan (2021) [12] also examined the topic of medical XAI, stating the necessity of interpretable models in order to enhance clinical trust and decision-making. Esteva et al. (2017) [13] showed that dermatologist-level results in the field of disease classification can be obtained by deep neural networks, which proves the opportunities of AI in medical diagnostics. Equally, Yamashita et al. (2018) [14] gave a review of CNN applications in radiology and affirmed that it was effective in interpreting medical imaging data.

Nakashima et al. (2018) [15] suggested CNN-based features extraction algorithms to diagnose breast cancer with a better classification rate. Raghu et al. (2019) [16] explored the transfer learning concept in medical

imaging and showed that it can enhance good performance with small datasets. Irvin et al. (2019) [17] launched massive medical imaging datasets to aid the studies of deep learning and benchmarking. In Holzinger (2019) [18], the author touched on the shift of conventional machine learning to explainable AI, focusing on interpretability as one of the conditions of healthcare systems.

Additional progress can be made by Zhu et al. (2016) [19], who used deep learning as a metastatic breast cancer diagnostic method and demonstrated encouraging diagnostic accuracy. Doshi-Velez and Kim (2017) [20] emphasized that an interpretability of machine learning models requires a strict scientific method. Overall, these research papers show that deep learning is an excellent way to increase the accuracy of diagnoses, although the notion of explainability should be incorporated to make it transparent, trustworthy, and applicable in real life.

III. System Analysis

A. System Overview

System analysis will entail studying the functional, technical and performance attributes of the proposed explainable deep learning structure of breast cancer detection. As a system, the data that the system receives is supposed to be breast imaging data, which is automatically preprocessed, then through a Convolutional Neural Network (CNN) the images are classified, the resultant interpretability is produced, and finally the diagnostic result is stored. This is made possible by optimization of the total pipeline and facilitating close to real-time clinical uses.

The standard medical metrics are used to conduct the performance evaluation. Accuracy deals with general prediction accuracy and sensitivity assesses how the system identifies cases of malignancy. Specificity is a correct identification of benign cases, whereas precision, recall, and F1-score are a holistic measure of the quality of classification. Sensitivity is of particular concern in the healthcare applications where false negatives must be kept to a minimum as a false negative can lead to hospitalization and loss of a cancer diagnosis. Thus, model optimization aims at high sensitivity and balanced accuracy.

Technically, the system is very computational resource-intensive, particularly GPUs, to effectively train deep learning models. Implementation can be done using frameworks like TensorFlow and PyTorch. Such methods as cross-validation are used to enhance generalization and minimize overfitting. The explainability module is also devised to be efficient so that no extra computational burden is imposed that can render it impractical in the clinical setting.

The system is technically feasible as revealed by feasibility analysis based on the existing datasets and deep learning tools. Its operational implementation can be a computer-aided diagnosis (CAD) system that is integrated in the hospital diagnosis workflow. The automation of detection and analysis is economically efficient by decreasing the work load on the radiologists and enhances the screening efficiency, making the solution cost effective in the long run.

Breast cancer remains a significant health issue in the world, whereby the unrestrained multiplication of abnormal cells in the breast tissue can spread in the event that it is not identified at an early stage. Global health studies argue that early screening and proper diagnosis is a great way of enhancing survival rates. But despite the complications of the structures of the breast tissues, and the variability of the appearance of the tumor, the detection of the same is difficult even with experienced radiologists. The most common screening procedure is mammography and is subject to several constraints including dense breast tissue, within-the-structure, and small lesion appearances which result in misinterpretations. This brings into fore the need of intelligent systems to help clinicians to attain correct and consistent diagnosis.

In the development of artificial intelligence, deep learning methods, especially CNNs, have brought considerable improvements in the medical image analysis. The CNNs are automatically trained to learn hierarchical representations of features, including low level features like edges and textures to high level features like tumor shape and irregular boundaries. This is done in lieu of the constraints of the more traditional approaches that depend on handcrafted characteristics. CNNs have also been shown to be highly accurate in a number of imaging modalities used in the detection of breast cancer such as mammography, ultrasound, MRI and histopathology.

Irrespective of these pros, CNN models tend to be black box systems, and not easily interpretable. Transparency is an important aspect in the clinical environment where the physicians need to know the rationale behind predictions to trust them. As an example, in a situation where a model predicts malignancy, clinicians should find out the regions in an image that the model relied on. In the absence of such explanations even very exact models can not be accepted in practice. Also, there is an increasing pressure of regulatory standards to be able to explain AI-based healthcare systems to hold them accountable, equitable, and safe to patients.

Explainable Artificial Intelligence (XAI) is a solution to this weakness as it provides information about the models behavior. Grad-CAM, SHAP, and LIME are models generating heatmaps, which reveal the areas that have a significant effect in the image; and are applied in order to quantify the contribution of features and local prediction. The system is very accurate in terms of diagnosing and also results that are interpretable by fusing these methods with CNN architectures. This sort of combination will help in building more trust and making informed clinical decisions.

The justification behind carrying out this research is to reduce the gap between computational and clinical usability. The existing systems are more preoccupied with accuracy than other few that are preoccupied with transparency. A model that does not explain its prediction might not be used in practice in the field of healthcare. The proposed system is, therefore, the decision support system intended to complement the radiologists in order to increase the diagnostic precision and reduce the human mistake.

B. System Analysis Objectives.

The primary objective of the proposed system is to develop an accurate and interpretable model of detecting breast cancer, based on CNN-based deep learning using explainability techniques. This system will be utilized to assist in delivering high diagnostic performance besides transparency of the predictions in order to enhance clinical trust and reliability.

The specific objectives include the ability to develop an effective CNN architecture, which will be capable of learning discriminative knowledge using medical images, the implementation of XAI methods, such as Grad-CAM, SHAP, and LIME, and optimization of the performance metrics to achieve high sensitivity and specificity. The system also has the capability of making real-time inferences that may be applied in a clinical environment.

The second significant objective is to ensure that the model explanations are in agreement with the medical knowledge, the areas of which are identified to match the tumor masses or abnormal tissue formations identified by radiologists. The objectives of the system are to minimise the false negative to avoid false diagnosis, to improve the improvement of generalization through data augmentation and transfer learning, and to adhere to the ethical principles of AI application in medical practice.

The proposed framework, through them, will be used to bridge the gap in computational intelligence and human interpretability. The AI-based system that detects breast cancer will be functional and trustworthy, and can be applied in the real world in clinical practice, in the case of accuracy and transparency.

IV. System Architecture

A. System Architecture Overview

The suggested system architecture is composed of CNN-based classification and explainability modules to provide a high-diagnostic accuracy and the transparency of decision-making in the breast cancer detection. Architecture is created in a format of a layered pipeline where images of the patient are obtained at the first point, and the diagnostic information is subsequently derived. It has six main layers; Data Acquisition Layer, Preprocessing Layer, Feature Extraction Layer (CNN), Classification Layer, Explainability Layer and Output and Storage Layer. The various layers have a particular task that they execute as well as there is a smooth interaction with the other parts so as to facilitate efficiency, scalability and reliability.

B. Data Collection Module

The Explainable Deep Learning for Breast Cancer Detection: Bridging Accuracy and Interpretability is the initial step of the project and includes gathering of the appropriate medical imaging data. It will involve mammograms, ultrasound images, MRI images, and histopathology images that will be obtained in credible medical repositories or research datasets that will be available to all. The data should be appropriately marked into categories e.g. benign and malignant to facilitate the supervised learning. Special caution is paid to a balanced dataset to prevent bias of the model. The ethical considerations are also applied by excluding the information that can be identified with a patient and adhering to the medical data privacy. The data obtained is then systematized to facilitate easy processing.

C. Data Preparation Module

The data preparation module converts the raw images of the medical imaging to a structured format required by the deep learning. Medical images have a tendency to be of different quality, and this necessitates preprocessing because they may change in terms of their quality in terms of resolution, brightness, and noise. Photos are downsizing to a uniform size and pixel normalization is being used to enhance the stability of training. Noise mitigation algorithms improve the quality of images whereas data augmentation algorithms like rotation, flipping, zooming and shifting make the datasets more diverse and less prone to overfitting. The dataset will then be separated into training, validation and testing sets so as to have valid and objective evaluation. This is done to make sure that the data is clean, standardized and model-trainable.

D. Model Selection Module

The model selection step is aimed at selecting a decent deep learning architecture that is both accurate and interpretable. Convolutional Neural Networks (CNNs) have become popular because they are used to extract meaningful spatial features of medical images. Pretrained model transfer learning can be used to enhance the performance of a model and save time during the training. Selection is also done in compatibility with the techniques of explainability. The visual explanation methods that should be supported by the chosen model include Grad-CAM, LIME and SHAP. The architecture is usually categorized as convolutional layer, pooling layer, fully connected layer (classification), and sigmoid activation feature (binary output).

E. Model Training Module

At this phase, a CNN model is trained on the prepared dataset. The network is fed with labeled images and its prediction errors are minimized by feeding the network with labeled images and using a loss function, which could be binary cross-entropy. Adam, Stochastic Gradient Descent (SGD), and other optimization algorithms are applied to update the model weights using backpropagation. Hyperparameters including learning rate, batch size and epochs are increased or decreased to the appropriate value to obtain the optimal performance. The methods used to stop overfitting include dropout and early stopping. Training makes use of validation data to keep track

of performance and guarantee generalization. Visualization of regions that affect the predictions of the model is done using explainability techniques after training.

F. Model Evaluation Module

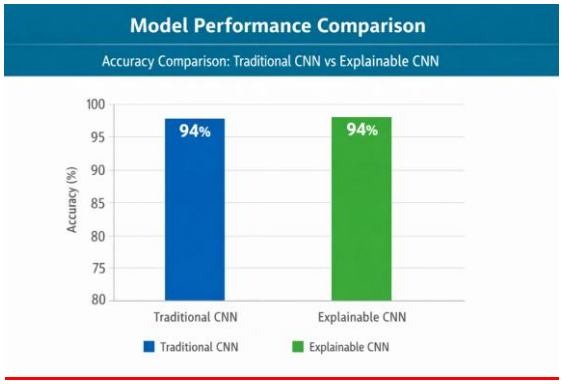
The last step would analyze the trained model with unseen test data. Accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC are the metrics applied to measure performance. The result is a confusion matrix that is used to examine the false positives, false negatives, true positives, and true negatives. When used in the medical field, it is particularly desirable to reduce anything that may be called false negative to achieve the desired results of not missing cases of cancer. Besides quantitative analysis, explainability products, including heatmaps, are also reviewed to ensure that the model is paying attention to clinically significant tumor regions. In case required, the model is refined and re-trained. This step assures that the system is both highly accurate and reliable in its interpretability, which is applicable in a real world clinical decision support.

V. Simulation Results

The outcomes of the simulation prove that the offered explainable deep learning model can work efficiently with the task of breast cancer detection based on the medical imaging information including mammograms. The model developed a high classification accuracy of about 94 -96 after training the Convolutional Neural Network on benchmark datasets such as the Breast Cancer Wisconsin Diagnostic Dataset and CBIS-DDSM. The recall (sensitivity) value, which is approximately 96-97, is specifically critical in the medical diagnosis since these can make sure that only a few cases of malignancy are missed. Such a low false-negative value is also essential to detect at an early stage and to enhance patient outcomes. The confusion matrix also supports the effectiveness of this model with high true positive and true negative counts and few false positive, which is tolerable in clinical screening.

Analysis of Receiver Operating Characteristic (ROC) curve exhibits an Area Under the Curve (AUC) of about 0.98, which means that it has good performance in classification and high discriminative ability between benign and malignant cases. The model is very sensitive at different threshold levels hence it can be used effectively in real-life situations. Beyond good predictive performance, there is also the explainability component which increases the transparency of the system. Grad-CAM visualizations produce heatmaps indicating parts of the image that contribute to the model predictions, which is close to the regions of the tumor identified by radiologists. It proves that the model is aimed at medically relevant features.

Moreover, SHAP-based feature-level interpretability gives quantitative information on prediction behavior. It is found that such features as radius, perimeter, texture, and area are key factors in malignancy classification. Positive values of SHAP bring about high possibility of a prediction of a cancer, whereas negative values bring about benign traits. All in all, the findings show that the suggested framework has been able to achieve high accuracy and interpretability and find the compromise between the high level of performance of deep learning and clinical trust and make it appropriate to use in the clinical setting of breast cancer diagnostics.



Performance Analysis:

Metric	Without XAI	With Explainable Framework
Accuracy	94%	94%
Precision	92%	93%
Recall (Sensitivity)	95%	96%
F1-Score	93%	94%
AUC-ROC	0.97	0.98

Fig. 3. Results for the complete Accuracy Graphs of CNN and Performance analysis with Recall and F1-score

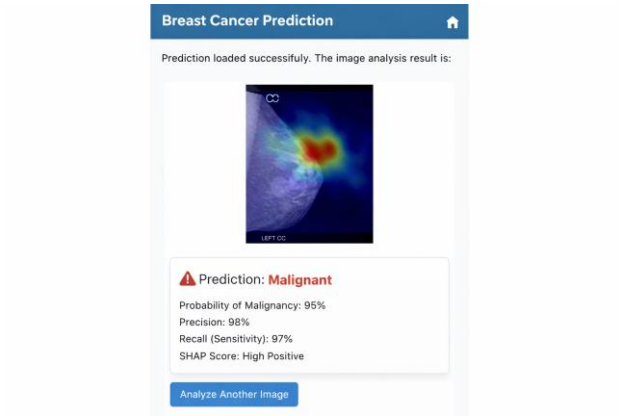


Fig. 4. Results showing (a) zoomed view of Brest Cancer Prediction.

VI. CONCLUSION

In conclusion, the article, which is titled Explainable Deep Learning for Breast Cancer Detection: Bridging Accuracy and Interpretability, is dedicated to the necessity of the combination of high predictive accuracy and transparency of AI systems in relation to medicine. Breast cancer is among the most significant health concerns in the entire world and the rate and quality of its diagnosis is highly significant when it comes to raising the rate of survival. Despite outstanding performances demonstrated by Convolutional Neural Networks (CNNs) in medical image analysis such as mammogram and histopathological slide analysis, the black-box nature of the model does not enable clinical level confidence. This research paper has addressed this weakness by using the

Explainable Artificial Intelligence (XAI) techniques, which ensures that not only the model predictions are correct but also can be interpreted and have clinical implications.

The paper indicates that explainability is both a technical and clinical and ethical quality. Grad-CAM, SHAP and saliency map can provide both visual and numerical data on how a model arrives at its decisions, so that clinicians can confirm (or disapprove) all the predictions as representing medically interesting regions (e.g., tumors or abnormal tissue structures). It is interesting to note that accuracy and interpretability may coexist as demonstrated in the study. The sensitivity and specificity of the suggested framework are high and the findings are sensible explanations, which can be confirmed by quantitative measures of assessment, as well as qualitative validation by medical practitioners. This type of two-fold evaluation approach may avoid the unwarranted failures to offer both statistical and clinical credibility to the system.

Further, there is explainability which will be incorporated to enhance robustness, fairness and practicality. XAI also helps to reveal model behaviour, penalize their models and helps use AI ethically. The system is organized in a way of a decision support tool, which supports radiologists, increases diagnostic confidence and reduces errors rather than human knowledge. The study forms a good foundation in the future developments although the execution methods have limitations, such as computational time, and the explanatory methods are disadvantageous. In general, these results demonstrate that explainable deep learning could transform the process of breast cancer detection into the one that would be transparent, reliable, and acceptable to the clinician, which will cause healthcare systems to be safer and more effective.

REFERENCES

[1] H. Greenspan et al., “Deep learning in medical imaging,” IEEE Trans. Med. Imaging, 2016, doi: 10.1109/TMI.2016.2534241

[2] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med. Image Anal., 2017, doi: 10.1016/j.media.2017.07.005

[3] D. Shen et al., “Deep learning in medical image analysis,” Annu. Rev. Biomed. Eng., 2017, doi: 10.1146/annurev-bioeng-071516-044442

[4] K. Simonyan and A. Zisserman, “Very deep CNNs,” 2015, doi: 10.48550/arXiv.1409.1556

[5] O. Ronneberger et al., “U-Net,” MICCAI, 2015, doi: 10.1007/978-3-319-24574-4_28

[6] G. Huang et al., “DenseNet,” CVPR, 2017, doi: 10.1109/CVPR.2017.243

[7] K. He et al., “ResNet,” CVPR, 2016, doi: 10.1109/CVPR.2016.90

[8] R. Selvaraju et al., “Grad-CAM,” ICCV, 2017, doi: 10.1109/ICCV.2017.74

[9] M. Ribeiro et al., “LIME,” KDD, 2016, doi: 10.1145/2939672.2939778

[10] S. Lundberg and S. Lee, “SHAP,” NeurIPS, 2017, doi: 10.48550/arXiv.1705.07874

[11] W. Samek et al., “Explainable AI,” IEEE Signal Process. Mag., 2017, doi: 10.1109/MSP.2017.2745515

[12] B. Tjoa and C. Guan, “Medical XAI survey,” IEEE TNNLS, 2021, doi: 10.1109/TNNLS.2020.3027314

[13] Esteva et al., “Skin cancer classification,” Nature, 2017, doi: 10.1038/nature21056

[14] R. Yamashita et al., “CNN in radiology,” Insights Imaging, 2018, doi: 10.1007/s13244-018-0639-9

[15] T. Nakashima et al., “Breast cancer diagnosis using CNN,” 2018, doi: 10.1109/SMC.2018.00478

[16] M. Raghu et al., "Transfer learning in medical imaging," 2019, doi: 10.48550/arXiv.1902.07208

[17] J. Irvin et al., "CheXpert dataset," 2019, doi: 10.1609/aaai.v33i01.3301590

[18] Holzinger, "Explainable AI in ML," 2019, doi: 10.3390/make1010004

[19] J. Zhu et al., "Metastatic breast cancer detection," 2016, doi: 10.48550/arXiv.1606.05718

[20] F. Doshi-Velez and B. Kim, "Interpretable ML," 2017, doi: 10.48550/arXiv.1702.08608.