International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Classification of Breast Cancer Using Machine Learning Models

Naisha E Shelke*, Dr. Ruhin Kouser**

*(Department Computer Science and Engineering, Presidency University, Bengaluru, India
Email: naishashelke@gmail.com)

** (Department Computer Science and Engineering, Presidency University, Bengaluru, India
Email : ruhinkouser@presidencyuniversity.in)

----------**********----------

Abstract:

One of the most prevalent forms of cancers among women globally and a leading cause of cancer-related deaths. This paper discusses the use of hybrid machine learning models combined with VGG16 for the identification of breast cancer subtypes and normal tissues. Four models used in this work are RF-VGG16, XGBoost-VGG16, SGD-LOG-VGG16, and LightGBM-VGG16, respectively, to compare their performance in accurately identifying invasive ductal carcinoma, invasive lobular carcinoma, mucinous carcinoma, and normal tissues using a dataset of 1,016 mammogram images. Among these models, LightGBM-VGG16 yields the best classification accuracy of up to 88%. The recall was balanced among the majority of the classes for each model while indicating challenges such as feature overlap and class imbalance. These results highlight the potential of hybrid models in enhancing early detection of breast cancer while underscoring the need for better data representation and advanced machine learning techniques to achieve more reliable diagnostic outcomes.

Keywords — **Breast Cancer, Classification, Machine Learning, Hybrid models, Accuracy, Precision.**

----------**********----------

I. INTRODUCTION

Breast cancer is one of the most prevalent and significant health challenges globally, particularly among women. Ac- counting for 2.3 million new cases and approximately 685,000 deaths in 2020 alone [1], it has become the leading cause of cancer-related mortality in women. Breast cancer develops primarily in the milk ducts or lobules, which are responsible for milk production. Early detection is crucial because it directly correlates with improved survival rates and better treatment outcomes. Despite advancements in treatment methods—such as surgery, radiation, chemotherapy, immunotherapy, and targeted therapies—the disease's high prevalence and mortality rates underscore the necessity for continuous improvements in diagnostic approaches. In recent years, medical imaging technologies such as mammography, ultrasound, and magnetic resonance imaging (MRI) have been instrumental in breast cancer diagnosis. However, accurately interpreting these imaging results remains a challenge, often leading to delayed diagnosis or false positives and negatives. Machine learning and deep learning techniques have revolutionized breast cancer detection and classification by providing automated, efficient, and accurate diagnostic tools. These techniques analyse vast amounts of imaging data to detect patterns and anomalies that are difficult for humans to identify. Specifically, hybrid models that integrate deep learning architectures like VGG16 with advanced machine learning algorithms hold immense potential in improving the accuracy and reliability of breast cancer diagnosis. These models leverage the feature extraction capabilities of deep convolutional neural networks while addressing challenges such as feature overlap, class imbalance, and data representation issues. This study investigates the application of hybrid machine

learning models for the classification of breast cancer subtypes and normal tissues, focusing on invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), mucinous carcinoma, and normal tissues. By comparing the performance of various models, including RF-VGG16, XGBoost-VGG16, SGD-LOG-VGG16, and LightGBM-VGG16, this research aims to highlight the strengths and limitations of different approaches in breast cancer diagnosis. Furthermore, this work emphasizes the critical role of data preprocessing, feature engineering, and model optimization in achieving reliable diagnostic outcomes.

II. ABOUT CLASSES

Breast cancer encompasses a range of histological subtypes, each with distinct characteristics and implications for diagnosis and treatment. Invasive ductal carcinoma (IDC) is the most prevalent subtype, accounting for approximately 80% of all cases. It begins in the milk ducts and invades surrounding breast tissues, often leading to the formation of palpable lumps. Invasive lobular carcinoma (ILC), which represents about 10-15% of cases, originates in the milk-producing lobules and tends to spread diffusely, making it harder to detect during physical examinations. Mucinous carcinoma is a rarer subtype characterized by the production of mucus within the tumour, offering a relatively favourable prognosis compared to IDC and ILC. Lastly, normal tissue samples are included in this study as a baseline for comparison, aiding in the differentiation of cancerous and non-cancerous cases. Understanding strategies and patient outcomes.

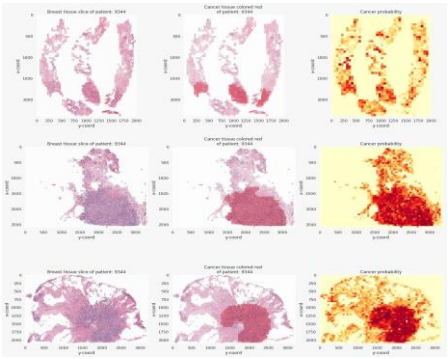

FIG 1.

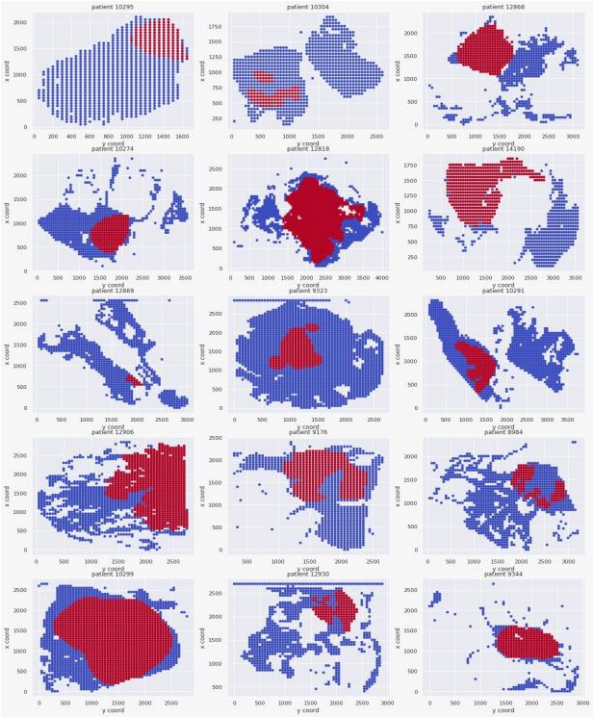

FIG 2.

III. LITERATURE REVIEW

The application of machine learning in breast cancer detection has been widely studied, showcasing significant advancements in diagnostic accuracy and early detection. Lakshmanaprabu et al. proposed an Optimal Deep Neural Network- Linear Discriminant Analysis (ODNN-LDA) model for mammogram classification, which reduced feature dimensionality and enhanced classification performance, achieving high ac- curacy in distinguishing benign from malignant tumours [2].

Abid et al. introduced a multi-view Convolutional Recur- rent Neural Network (MV-CRecNet) for breast cancer sub- type classification using mammogram datasets. This approach leveraged multiple perspectives of mammograms to achieve accuracies of 96.8% and 98.5% on different datasets, demonstrating the effectiveness of deep learning in medical imaging [3].

Tiwari et al. explored the use of transfer learning with Convolutional Neural Networks (CNNs) for breast cancer detection. Their study achieved remarkable results by fine- tuning pre-trained

models such as ResNet and InceptionV3 on mammogram datasets, achieving up to 95% accuracy [4]. Zheng et al. proposed a novel hybrid approach combining Support Vector Machines (SVMs) with deep learning architectures to classify breast cancer subtypes[5].

Jena et al. analysed the performance of multiple deep learning models for mammogram classification, highlighting the importance of data augmentation and class balancing techniques in improving diagnostic outcomes. Their study pro- vided insights into optimizing model architectures for medical image analysis [6].

Martinez et al. developed a feature extraction and classification framework using artificial intelligence.Their approach combined handcrafted features with deep learning models to achieve a comprehensive analysis, yielding an accuracy of 93% [7].

These studies underscore the growing potential of hybrid machine learning approaches in breast cancer detection, emphasizing the need for further research into advanced data pre- processing techniques, model optimization, and the integration of multi-modal datasets to enhance diagnostic reliability.

A. Dataset

The dataset used in this study was sourced from Kaggle and consists of 1,016 images, which were split into training, validation, and test sets. These images represent four categories: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal cells. The dataset provides a varied collection of labelled images, enabling the development and assessment of machine learning models for the detection of lung cancer.

Several advanced machine learning models for image classification have been explored in this research, leveraging the feature extraction capabilities of VGG16 CNN with various classification algorithms for breast cancer detection. The RF VGG16 approach combines the Random Forest algorithm, which constructs an ensemble of decision trees to classify im- ages based on features extracted by VGG16. Random Forest's advantages, such as resilience to overfitting and its ability to handle noisy data, complement the rich image features learned by VGG16 effectively.

XGBoost VGG16 employs the Extreme Gradient Boosting algorithm, a powerful ensemble learning method that builds decision trees sequentially to correct prior errors, providing high predictive accuracy. Here, VGG16 performs feature ex- traction, while XGBoost fine-tunes the classification process. SGD Log VGG16 integrates stochastic gradient descent optimization with logistic regression. The deep features extracted by the VGG16 model are further optimized using stochastic gradient descent, enabling logistic regression to achieve an optimal solution in shorter time intervals, making it efficient for breast cancer image classification.

Lastly, LightGBM VGG16 combines VGG16's deep feature extraction with the LightGBM framework, a gradient boosting method known for its efficiency in speed and memory usage. LightGBM develops decision trees for classification, using the rich feature set learned by VGG16 to achieve accurate and efficient results, especially when working with large datasets. In all cases, the powerful deep learning capabilities of VGG16 are paired with advanced machine learning algorithms, resulting in a hybrid approach that harnesses the strengths of both deep learning and traditional machine learning methods. This enhances the precision and reliability of breast cancer classification.

IV. MACHINE LEARING ALGORITHMS

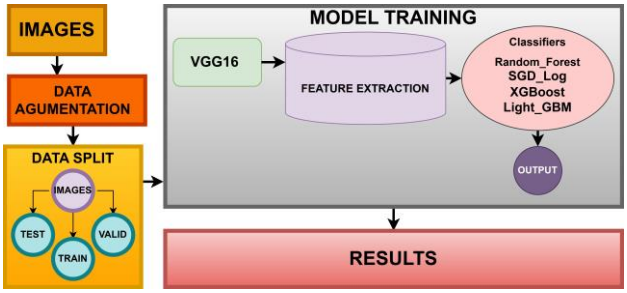

FIG 3. Flowchart of Classification of Breast Cancer using Machine Learning Algorithms

V. RESULTS

The results of the RF-VGG16 model, which reveals variations in performance across different lung cancer classes. The classification report, where the model excels at classifying adenocarcinoma, achieving a high recall of 0.97, though the precision is lower at 0.62, with an F1-score of 0.76. Large cell carcinoma follows, with a recall of 0.47 and an F1- score of 0.62, demonstrating that while the model detects most cases, it struggles to find all instances of this class. Squamous cell carcinoma achieves a precision of 0.92 but has a lower recall of 0.49 and an F1-score of 0.64. The model is perfect at classifying normal tissue, with a precision of 1.00, a recall of 0.98, and an F1-score of 0.99. Overall, the model's accuracy is 75%. Figure 3 further elaborates on the model's performance through a confusion matrix, showing that adeno- carcinoma is mostly well-classified, while large cell carcinoma and squamous cell carcinoma are frequently misclassified, often as adenocarcinoma. The feature importance plot from the Random Forest method highlights which features were most influential in the classification process. These results suggest that, while the model performs well with some classes, particularly normal tissue, improvements are needed in recall for large cell carcinoma and squamous cell carcinoma to enhance overall classification accuracy.

XGBoost-VGG16 model in classifying breast cancer types and normal samples. The classification reports an overall accuracy of 83%, with a macro average F1-score of 0.84, reflecting the model's balanced performance across all classes. Precision is highest in the Normal class, where no false positives are reported, achieving a perfect score of 1.00, although the recall is slightly lower at 0.96, indicating some missed cases. The F1-score for adenocarcinoma is 0.83, with a high recall of 0.93, suggesting the model identifies most adenocarcinoma instances with few false negatives. In contrast, large cell carcinoma and squamous cell carcinoma have lower recall values of 0.69 and 0.71, respectively, reflecting the model's difficulty in detecting all instances of these types. These challenges further through a confusion matrix, showing that most samples are correctly classified, but significant misclassifications occur between adenocarcinoma and squamous cell carcinoma. The feature importance plot reveals that certain critical features extracted by VGG16 play a significant role in classification decisions. The model performed well for both normal tissue and adenocarcinoma cases but needs further optimization, particularly to improve recall for the other cancer types, possibly due to feature overlap or class imbalance.

SGD LOG-VGG16 model. Figure 6 shows an overall ac- curacy of 63%, with a macro average F1-score of 0.60 and a weighted average F1-score of 0.55. The model performs well in classifying normal cases, with a precision of 1.00, a recall of 0.93, and an F1-score of 0.96, indicating strong accuracy for this class. For the "large cell carcinoma left hilum T2 N2 M0 IIIa" class, precision is 0.96, but recall is lower at 0.53, pointing to misclassification issues. The poorest performance is observed for "squamous cell carcinoma left hilum T1 N2 M0 IIIa," which has a recall of 0.03 and an F1-score of 0.06, likely due to class imbalance and feature misrepresentation. Confusion matrix represents the most "squamous cell carcinoma" samples are misclassified as "adenocarcinoma." Although adenocarcinoma has a high recall of 0.99, its precision is relatively low at 0.51, suggesting overprediction for this class. The feature coefficients plot indicates some model overfitting to certain features. However, the correct

prediction for a normal case in bottom-right corner validates the model's reliability for this class. These results highlight the model's strength in handling well-represented classes like normal tissue but indicate the need for optimization to address underrepresented classes like squamous cell carcinoma.

LightGBM-VGG16 model, which combines the LGBM classifier with a VGG16-based CNN for classification of adenocarcinoma, large cell carcinoma, normal tissue, and squamous cell carcinoma. The classification report shows an average accuracy of approximately 86% across all samples, with a macro average F1-score of 0.87, reflecting balanced performance across all image classes. Precision, recall, and F1- scores are generally good, with the "normal" class achieving an F1-score of 0.98 and "large cell carcinoma" achieving an F1-score of 0.84, though with somewhat lower recall values of 0.76 and 0.77, respectively. The confusion matrix in Figure 9 shows that the majority of predictions are correct, with only minor misclassifications, such as 21 instances of squamous cell carcinoma being incorrectly identified as adenocarcinoma. The feature importance plot indicates that a few key features have a significant impact on the model's predictions, while the rest contribute less. Additionally, a CT scan visualization of a correctly classified "normal" case further demonstrates the model's reliability when processing medical images. These results confirm the LightGBM-VGG16 model's effectiveness and robustness in classifying complex medical data.

The comparison of the RF-VGG16, XGBoost-VGG16, SGD LOG-VGG16, and LightGBM-VGG16 models reveals that the LightGBM-VGG16 model, with an accuracy of 86%, provides the most balanced performance across all classes. While RF-VGG16 and XGBoost-VGG16 also show strong results, both face challenges with recall. Feature optimization and addressing class imbalances are crucial steps for improving classification performance in all models.

VI. DISCUSSION

The findings of this research highlight both the potential and challenges of applying machine learning models, particularly those incorporating feature extraction from VGG16, for classifying breast cancer subtypes and normal tissues. Four different hybrid models were evaluated: RF-VGG16, XGBoost-VGG16, SGD-LOG-VGG16, and LightGBM-VGG16, which demonstrated varying strengths and weaknesses in handling the complexities of breast cancer classification from medical images. Among these, the LightGBM-VGG16 model achieved the most balanced performance, with an accuracy of 86%, as well as high precision and recall for most of the classes. How- ever, it struggled to provide accurate results for some specific subtypes of breast cancer, posing significant challenges and indicating the need for further optimization.

The overall accuracy of RF-VGG16 was moderate at 75%, but it showed considerable variation across different classes. While it was efficient at identifying normal tissue and some breast cancer subtypes, it faced challenges distinguishing be- tween certain subtypes. Its low recall values suggest issues related to class imbalance or feature overlap, making it difficult to effectively differentiate between some of the breast cancer types. Random Forest was resilient to overfitting and noise, leveraging VGG16's ability to extract features, but its reliance on ensemble decision trees limited its ability to detect subtle differences between certain cancer subtypes.

The XGBoost-VGG16 model performed better, achieving an accuracy of 83% with a macro-average F1-score of 0.84. It showed strong performance in classifying normal tissue and certain breast cancer types, but like RF-VGG16, it struggled with some subtypes. XGBoost, being a sequential algorithm, was able to correct its mistakes in subsequent rounds, helping improve performance. However, misclassifications between specific subtypes highlighted that even though VGG16 ex-

tracts rich features, they may not fully capture the unique characteristics of these subtypes. This suggests the need for advanced feature engineering or data augmentation techniques to improve the feature representation for these challenging cases.

The SGD-LOG-VGG16 model was the least effective, with an accuracy of 63% and a macro-average F1-score of 0.60. Despite accurately classifying normal tissue with high precision, its performance for other breast cancer subtypes was subpar, particularly for the more challenging cases, where the recall for some subtypes was very low. This indicates significant issues with class imbalance and the model's sensitivity to underrepresented subtypes. The simple nature of logistic regression combined with VGG16's powerful feature extraction was not sufficient to tackle the complexities of the dataset, emphasizing the need for more robust optimization strategies or better-balanced datasets.

Finally, LightGBM-VGG16 model outperformed the others, with the highest accuracy of 86%. It demonstrated balanced precision, recall, and F1-scores across most classes, performing well in terms of handling large datasets and creating deep decision trees. However, misclassifications be- tween certain subtypes, such as the overlap between some breast cancer types, presented challenges. To improve this, further efforts in data preprocessing, such as addressing class imbalance through oversampling, under sampling, or synthetic data generation, are needed.

In conclusion, while the study showcases the promise of using hybrid VGG16 models for classifying breast cancer, the results also highlight ongoing challenges such as class imbalance, feature overlap, and the need for optimization. Future work should focus on improving feature representation, utilizing more diverse and balanced datasets, and exploring more advanced machine learning techniques to enhance diagnostic accuracy and reliability.

VII. CONCLUSION

In conclusion, this research demonstrates the potential of hybrid machine learning models that combine VGG16 feature extraction with algorithms like Random Forest, XGBoost, SGD-LOG, and LightGBM for the classification of breast cancer subtypes and normal tissue. Among these models,

LightGBM-VGG16 achieved the best performance with an accuracy of 86%, showing balanced precision and recall across most classes. However, challenges such as misclassifications between certain subtypes and the need for further optimization were evident, particularly in handling class imbalance and feature overlap.

The results highlight the importance of addressing these issues through advanced feature engineering, class balancing techniques, and more sophisticated models. While RF-VGG16 and XGBoost-VGG16 performed well in some areas, they struggled with certain subtypes due to low recall and feature limitations. The SGD-LOG-VGG16 model, while strong in classifying normal tissue, exhibited poor performance for less represented classes, indicating the need for improved data representation.

Overall, this study underscores the promise of integrating deep learning with traditional machine learning techniques for medical image classification but also emphasizes the need for further refinement to improve the accuracy and robust- ness of models, particularly in challenging cases. Future re- search should focus on enhancing feature extraction, balancing datasets, and optimizing machine learning models to achieve more reliable and accurate breast cancer diagnoses.

REFERENCES

[1] Simonyan, K., & Zisserman, A. (2015). Very deep convolu- tional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.
[3] Sirinukunwattana, K., Raza, S. E. A., & et al. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5), 1196-1206.
[4] Khan, A., & et al. (2020). Breast cancer classification using machine learning algorithms and deep learning networks: A review. *IEEE Access*, 8, 70191-70201.
[5] Chou, W., & et al. (2020). A novel hybrid deep learning model for breast cancer detection. *Journal of Healthcare Engineering*, 2020.
[6] Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumo- nia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
[7] Xie, Y., & et al. (2018). The role of machine learning algorithms in breast cancer diagnosis: A review. *Artificial Intelligence in Medicine*, 95, 43-59.
[8] Zhou, X., & et al. (2020). Breast cancer histopathological image classification using convolutional neural networks. *Computers in Biology and Medicine*, 122, 103801.
[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting sys- tem. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[10] Ke, G., Meng, Q., & et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

