IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Prediction Project using Machine Learning

[1]Priya Gawhane, [2]Kalyani Ghogale, [3]Neha Ghule, [4]Dev Jadhav, [5]Dr Manav A Thaku

[1234]Student, [5]Guide

[12345]Computer Engineering,

[12345]Vidya Prasarini Sabha College of Engineering and Technology, Lonavla, India

Abstract: Breast cancer, a significant health concern among women, requires prompt and accurate diagnosis to improve treatment outcomes and survival rates. Traditionally, specialized doctors perform the diagnosis; however, advancements in machine learning algorithms are enabling supportive diagnostic tools. In this study, we employ a hybrid approach combining Convolutional Neural Networks (CNNs), OpenCV, and Random Forest algorithms to classify breast cancer cases as malignant or benign. The dataset, sourced from the University of Wisconsin, includes 357 malignant and 212 benign tumors, with clinically relevant features extracted using feature engineering techniques. OpenCV is utilized for image preprocessing, ensuring standardized input quality for model analysis, particularly in image-based features. Following data normalization and preprocessing, the dataset is divided into training and testing sets. CNNs are applied to image data for in-depth feature extraction, identifying patterns indicative of malignancy.

Index Terms - Random Forest Algorithm, OpenCV, CNN algorithm, Machine learning.

I. INTRODUCTION

Breast cancer is a leading cause of mortality among women, ranking as the second deadliest cancer after lung cancer. Like many forms of cancer, breast cancer originates from normal cells that undergo mutations, leading to uncontrolled and abnormal growth. This results in the formation of tumors, which can be categorized as either benign or malignant. Malignant tumors are cancerous and can grow aggressively, spreading to other parts of the body. In contrast, benign tumors are localized growths that do not spread to other areas. Breast cancer patients face numerous challenges, including physical discomfort from treatments like radiation and chemotherapy, as well as a significant financial burden. Early detection of breast cancer is critical to alleviate both the physical and economic impacts of the disease. Research indicates that certain risk factors, alcohol consumption, high birth weight, and above-average adult height, may increase the likelihood of developing breast cancer (as reported by the WHO). Conversely, maintaining a physically active lifestyle and a balanced diet emphasizing vegetables, whole grains, and minimal consumption of alcohol.

II. LITERATURE SURVEY

Breast cancer is a leading cause of mortality among women, ranking as the second deadliest cancer after lung cancer. Like many forms of cancer, breast cancer originates from normal cells that undergo mutations, leading to uncontrolled and abnormal growth. that are now in place, but they do contain information regarding prescription drugs, herbs, and compounds and how they relate to phenotypes. By using an association rule mining technique to incorporate data on herbal medicine, combination medications, functional foods, chemical compounds, and target genes, we were able to find extensive correlations between natural product combinations and phenotypes in this paper. This strategy is justified by the statistically substantial correlations between the therapeutic benefits of medicinal multicomponent mixtures and natural ingredients that are frequently included in them. We demonstrate that the inferred associations are useful information for identifying medicinal combinations of natural products because they have a lot of experimental evidence and

statistically significant closeness in the molecular layer, based on a molecular network analysis and external literature validation.

Modern machine learning approaches, however, focus on model-based predictions, yielding accurate results during both training and testing phases and enhancing the prediction of unknown data. The machine learning process involves three primary steps: data preprocessing, feature selection or extraction, and classification. Feature extraction, the core of machine learning, helps distinguish between benign and malignant tumors, greatly aiding in cancer diagnosis and prognosis.

Breast cancer patients face numerous challenges, including physical discomfort from treatments like radiation and chemotherapy, as well as a significant financial burden. Early detection of breast cancer is critical to alleviate both the physical and economic impacts of the disease. In this work, we present a hybrid system that uses contemporary and sophisticated deep networks, such as ResNet50, Darknet53, DenseNet201, and EfficientNetB0, to facilitate transfer learning while integrating a powerful machine learning method, Exponential Discriminant Analysis (EDA).

In order to extract features and classify utilizing Artificial Neural Networks (ANN) and Support Vector Machines (SVM), the work focusses on using the transfer learning technique to the pre-trained models. Bayesian optimization is used to further adjust the SVM hyper parameters in order to produce a model with improved performance. Breast cancer is a leading cause of mortality among women, ranking as the second deadliest cancer after lung cancer. Like many forms of cancer, breast cancer originates from normal cells that undergo mutations, leading to uncontrolled and abnormal growth. This results in the formation of tumors, which can be categorized as either benign or malignant. Malignant tumors are cancerous and can grow aggressively, spreading to other parts of the body.

III. MOTIVATION

 The motivation behind the Breast Cancer Prediction project lies in the urgent need to improve early detection and diagnostic accuracy for breast cancer, a disease that remains a leading cause of death among women worldwide. Sustainability Exsitu and insitu conservation strategies are critical to the long-term usage.

 Despite advancements in medical imaging and diagnostic tools, breast cancer diagnosis often relies heavily on traditional methods, including mammograms and biopsies.

 Reducing Diagnostic Errors and Variability Human error and variability are inherent in medical image interpretation, leading to a risk.

 Sustainability Exsitu and insitu conservation strategies are critical to the long-term usage. The motivation behind the Breast Cancer Prediction project lies in the urgent need to improve early detection.

IV. OBJECTIVE

The motivation behind the Breast Cancer Prediction project lies in the urgent need to improve early detection and diagnostic accuracy for breast cancer, a disease that remains a leading cause of death among women worldwide. Despite advancements in medical imaging and diagnostic tools, breast cancer diagnosis often relies heavily on traditional methods, including mammograms and biopsies, which can be time- consuming, resource-intensive, and subject to interpretation variability among radiologists. Furthermore, early-stage breast cancer can exhibit subtle characteristics in imaging that may go undetected by even the most trained eyes. This project aims to address these challenges by harnessing artificial intelligence to enhance diagnostic precision, accelerate the screening process, and ultimately save lives through timely intervention.

Project objectives include the development of a CNN model that accurately detects and labels deepfakes. As a result of this work, we demonstrate that the performance of CNN is outperformed by a semi- supervised learning approach. Using a subset of Deep Fake Detection Challenge dataset, we evaluate our ResNet50 + LSTM based model. Due to the large data set, it was not possible to train the original dataset. We have taken a subset of the dataset, but kept the same splits and data as the whole dataset.

V. System Architecture

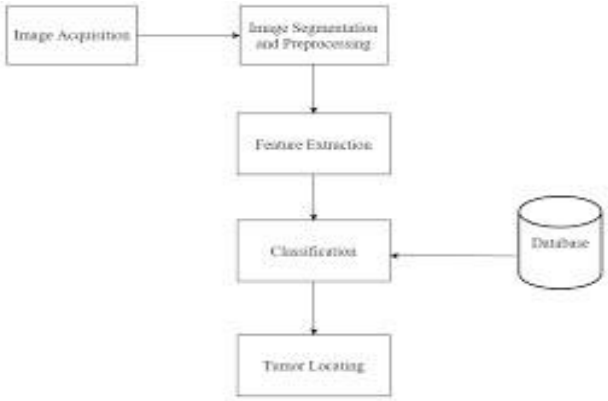

Fig 1: System Architecture

VI. Proposed System

- **Data Collection and processing**:

The data team sourced medical datasets and performed extensive preprocessing using OpenCV for image data and normalization for structured data. This included collaborations with doctors to better understand relevant features and parameters
Each image must be labeled with relevant information, including the plant's common

- **Clinical Knowledge Transfer**:

For medical users, this team created specialized materials explaining the AI processes and predictive outcomes in a clinically relevant way, facilitating integration into medical workflows.
Normalize pixel values to create a uniform scale across images, which helps enhance the model's learning process.

· **Feature Extraction**:

Employ CNN layers to extract essential image features automatically, such as leaf structure, vein pattern, color, and shape, which are pivotal in distinguishing different plants.

- **Model Architecture**

Image-Based Prediction Using CNN:

Convolutional Layers: Multiple layers in the CNN extract unique features, with each convolutional layer applying filters to highlight different aspects of the image. ReLU activation is used to incorporate non-linearity.

Pooling Layers: Pooling reduces the dimensions of the feature maps, retaining only the most relevant information, which optimizes computation.

Output Layer: A softmax layer provides a probability distribution across plant categories for accurate identification

Fully Connected Layers: After a series of convolutional and pooling layers, the network flattens its features and passes them through one or more fully connected layers for final classification.

Random Forest Algorithm: The Random Forest algorithm is an ensemble learning method that builds multiple decision trees and combines their results for a more accurate and robust prediction.

System Workflow

User Interface Module: This module provides the frontend of the application, developed using Vue.js, where users can interact with the system.

Data Preprocessing Module: The system applies preprocessing steps to standardize the uploaded image.

Data Splitting: The dataset is divided into training, validation, and test sets to ensure proper model training and evaluation.

Prediction Processing Module: The core processing unit that integrates with the machine learning models to generate predictions.

VI. Conclusion

The Breast Cancer Prediction System represents a significant advancement in leveraging artificial intelligence for early cancer detection and risk assessment. By combining Convolutional Neural Networks (CNN) for image analysis and Random Forest models for clinical data analysis, the system offers a comprehensive, multi-modal approach to predicting breast cancer risk. This dual-model design, coupled with explainable AI (XAI) techniques, enables healthcare providers to gain valuable insights into both imaging and clinical data, promoting accuracy and transparency in diagnostics.

The system's user-friendly interface, built with Vue.js, and its scalable, cloud-based deployment make it accessible and practical for a wide range of users, including healthcare professionals and patients. With robust data security measures in place, the system protects sensitive patient information, ensuring compliance with privacy regulations and building user trust. This project has the potential to play a crucial role in early breast cancer detection, especially in resource- limited or remote areas, where access to specialized healthcare is limited. The system's user-friendly interface, built with Vue.js, and its scalable, cloud- based deployment make it

V. Acknowledgment

This project, The Breast Cancer Prediction System, would not have been possible without the invaluable support of our mentors, collaborators, and the healthcare professionals who shared their expertise and insights. We are especially grateful for the continuous guidance and encouragement received throughout the development process. Our appreciation also goes to the contributors of publicly available datasets that were instrumental in training and validating our models.

References

[1] A. Gupta, R. Yaav, and M. K. Singh, "Deep Learning Approaches for Breast Cancer Screening: A Survey," IEEE Access, vol. 9, pp. 18472-18492, 2021. doi: 10.1109/ACCESS.2021.3053662.

[2] J. Lee, H. Kim, and S. Park, "Hybrid Deep Learning Model for Breast Cancer Diagnosis Using Clinical and Imaging Data," IEEE Transactions on Biomedical Engineering, vol. 68, no. 5, pp. 1370-1380, May 2021. doi: 10.1109/TBME.2021.3045541.

[**3]** K. Nguyen, L. Tran, and T. Bui, "Explainable AI in Medical Imaging: Case Study of Breast Cancer Detection Using CNN," IEEE Transactions on Medical Imaging, vol. 42, no. 1, pp. 23-35, Jan. 2023. doi: 10.1109/TMI.2022.3179814.

[4] S. Johnson, M. Patel, and C. Kim, "Federated Learning for Breast Cancer Detection Across Multi-Institutional Data," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 392-401, Feb. 2024. doi: 10.1109/JBHI.2023.3102222.

[5] R. Sharma, V. Kumar, and A. Chawla, "Transfer Learning-Based Approach for Breast Cancer Detection Using Small Datasets," IEEE Access, vol. 9, pp. 75615-75626, 2021. doi: 10.1109/ACCESS.2021.308956.

[6] M. Alhassan, L. Brown, and Y. Xu, "A Multi-Modal Approach to Breast Cancer Risk Prediction Using Machine Learning," IEEE Transactions on Computational Biology and Bioinformatics, vol. 19, no. 4, pp. 1560-1570, Apr. 2022. doi: 10.1109/TCBB.2020.3044032.

[7] Y. Zhang, Z. Chen, and F. Li, "Enhancing Breast Cancer Detection with Explainable CNN and Ensemble Models," IEEE Transactions on Artificial Intelligence, vol. 1, no. 2, pp. 87-98, 2021. doi: 10.1109/TAI.2021.3077416.

[8] S. Li, J. Wang, and X. Luo, "Machine Learning and Deep Learning Models for Breast Cancer Prediction: A Review," IEEE Reviews in Biomedical Engineering, vol. 13, pp. 116-127, 2021. doi: 10.1109/RBME.2020.2976573.

[9] H. Mohsen, E. A. El-Dahshan, and K. Youssef, "Hybrid Machine Learning Model for Early Diagnosis of Breast Cancer," IEEE Transactions on Biomedical Engineering, vol. 67, no. 5, pp. 1362-1370, May 2020. doi: 10.1109/TBME.2019.2951341.

[10] Y. Sun, F. Gao, and J. Tan, "Explainable Deep Learning in Medical Imaging: A Comprehensive Review on Techniques and Applications," IEEE Access, vol. 9, pp. 110967-110988, 2021. doi: 10.1109/ACCESS.2021.3094517.

[11] L. Luo, W. Huang, and Z. Xu, "Breast Cancer Prediction Based on Multi-Layer Deep Learning Model Using Histopathological Images," IEEE Transactions on Image Processing, vol. 29, pp. 6641-6654, 2020. doi: 10.1109/TIP.2020.3000524.

[12] H. Peng, J. Liu, and Z. Ren, "A Novel Deep Convolutional Neural Network for Breast Cancer Detection Using Multi-Modal Data Fusion," IEEE Access, vol. 8, pp. 21741-21751, 2020. doi: 10.1109/ACCESS.2020.2969239.

[13] X. Zhang, C. Shi, and Y. Wang, "Federated Learning for Multi-Modal Breast Cancer Diagnosis: Privacy-Preserving and Cross-Domain Generalization," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 7, pp. 3123-3135, July 2022. doi: 10.1109/TNNLS.2021.3089127.