BREAST CANCER DETECTION USING VGG19 AND HYBRID ML MODELS

M. Priyanka[1], K. Pavani [2], Sk. Nasrin[3]
#1 Assistant Professor, SRK Institute of Technology, Vijayawada

#2 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract

Breast cancer is one of the leading causes of death for women worldwide, and early detection is critical to increasing survival rates. However, making a diagnosis manually using ultrasound images takes a lot of time and can vary from one radiology to another. Using deep learning and ensemble machine learning methods, the goal of this study is to create an automated and dependable system for classifying breast cancer. For efficient feature extraction from ultrasound images, the proposed method makes use of a pre-trained VGG19 model. Image fusion with segmentation masks helps to highlight tumor regions. The extracted deep features are then classified using an Extra Trees Classifier (ETC), which leverages ensemble learning to improve prediction accuracy and generalization. To guarantee accurate evaluation, the dataset is preprocessed, normalized, and divided into training and testing sets. The results of the experiments show that the proposed VGG19 model with the ETC model outperforms conventional methods like Decision Tree and other baseline models in terms of accuracy, precision, recall, and F1-score. Additionally, the system has increased sensitivity and specificity, making it suitable for use in medical diagnosis. In conclusion, the combination of ensemble classification and deep feature extraction significantly improves the performance of breast cancer detection.

Keywords: *Breast Cancer Detection, Ultrasound Imaging, Deep Learning, VGG19, Extra Trees Classifier (ETC), Image Segmentation, Feature Extraction, Ensemble Learning, Medical Image Classification, Computer-Aided Diagnosis (CAD).*

I. INTRODUCTION

Breast cancer remains a major global health challenge, requiring accurate and timely diagnosis to improve patient survival rates. Because it is safe, cost-effective, and capable of detecting abnormalities in dense breast tissues, ultrasound imaging is frequently utilized for the purpose of screening for breast cancer. However, because manual ultrasound image interpretation is complicated and heavily reliant on the expertise of radiologists, it can result in diagnostic errors and inconsistencies. Automated systems that support medical image analysis with improved accuracy and efficiency have been made possible by recent advancements in artificial intelligence. Convolutional neural networks, in particular, as well as other deep learning models, have demonstrated impressive capacity for deriving meaningful features from medical images. By combining multiple decision models, ensemble learning methods have also demonstrated improved classification performance. A strong framework for accurate disease detection is created by combining these approaches. The use of image fusion techniques makes tumor regions even more visible, allowing models to focus on important areas more effectively. This research presents an intelligent system

that combines deep feature extraction and ensemble classification to achieve precise breast cancer detection. By reducing diagnostic workload and providing consistent and accurate predictions, the system is intended to assist medical professionals.

A. Objective

a intend is really to goal is to design an automatic ovarian cancer classification and it enhances treatment outcome as well as consistency utilizing advanced artificial intelligence technics. a approach that focuses through retrieving elevated showcases because after image data throughout a deep cnn as well as employing some kind ensemble classification of between operate rigorous categorization. a approach emphasises useful postprocessing, such as photograph generalization but also condensation to categorization hoods, to reinforce position includes after all investment. this facilitates effective managing yeah tall statistics but instead mitigates misclassified thru improved study skills. the target furthermore includes assessing this same framework employing appropriate performance measurement of between affirm the latter's efficiency throughout genuine situations but instead assistance medical judgement.

B. Problem Statement

diagnosis of breast cancer employing ultrasound provides challenges versus low light, sound, but also fluctuation throughout visual quality, that also hinder precise explanation. conventional testing systems based to either handbook appraisal, resulting in lack of consistency but also greater probability like human factor. prevailing deep learning nears frequently find it difficult to strong statistics but also lose the ability versus grasp intricate shapes successfully. new song classifier including such classification trees seem to be susceptible to clustering but instead underperform versus make generalisations well along data sets. is the need for a strong structure that really can proficiently selecta includes but also conduct accurate throughout various examples. attempting to address the above drawbacks has to some kind structural model that mixes computational intelligence such as extracting features as for ensemble learning such as managed to improve forecasting steady state but also accurateness.

C. Motivation

the expanding market regarding early and accurate tumor detection s driving any need for smart diagnostic methods the said assistance health provider. from either a business standpoint, computer machines decrease diagnosing time, lower operational costs, but also improve service effectiveness through health centers. from the a survey pragmatic point of view, assimilating computational intelligence to ensemble learning provides possibilities such as working to develop rising types able to handle complicated healthcare info. its advancements in medical image processing technology solutions much farther tries to encourage the event like credible ai-based options. the choice of such a concern would be pushed because of its genuine impact, potential to enhance clinical outcomes, as well as purview regarding technology through blending extraction of features but instead classification techniques.

II. LITERATURE REVIEW

1. muhammad abdel-nasser donc cetera. (2019): screening mammography through cancer screenings invariably continues to suffer that once inconstancies due to the variations with both numerous picturing perspectives, resulting in did miss as well as untrue detection systems. the target of all this investigate has been to enhance contains information related through color

coordinated tumour cell presidential contender notes all over distinct mammogram view points. a research methods means determining dubious provinces in every outlook but instead trying to establish communications among those utilizing corresponding methods to ensure conformance. through trying to integrate data from multiple view points, its system will reduce uncertainty but also enhances location yeah tumor cells. a test results demonstrated a major reducing through false - positive but also managed to improve algorithm to classify in comparison to single-view nears. the tactic elevates structural rigidity throughout recognizing cancer cell provinces through utilising geographic conformance all over pictures. this same work concluded and it number of co color coordinated offers one suitable solutions such as going to improve cancer detection system is a system and can also be successfully integrated in to other pc assessment structures, resulting in greater precise as well as clinically effective consequences.

2. d e. d e. cheng but instead f l. d u. modem (2017): cancer categorization utilizing neural network models frequently faces some challenges because of rigid teaching criteria the said restrict ability to adapt versus ranging available data. the target of all this project seeks to develop one identity multilayer perceptron infrastructure that really can change dynamically it's own learning. a research methods tends to involve trying to design of one neural as for responsive frameworks a certain enhance barbells but instead educational prices throughout going to train. the said approach allows a framework to reply efficaciously versus variants throughout data input as well as better key duties. the outcomes show greater precision as well as fast convergence compared with conventional dnns. its program displays excellent stability but also whittled down errors, especially for complex data points. a research concluded the said ego neurons provide us with a greater efficient and flexible workaround such as cancer categorisation, having contributed of between indicate significant reliability along biomedical imaging.

3. ahmed some one. berbar (2017): accurate after all breast and prostate cancer relies greatly upon that value yeah feature extraction, but instead single feature extraction techniques frequently fail to grasp full image features. the target of all this research would be to improve classification results using fusion techniques. its research methods combines feature selection method versus seize either of those mouth feel but instead structural data that once mammogram photographs. besides incorporating such characteristics, a system has created something a little more detailed depiction after all breast lesions. the outcomes demonstrate classifier accuracy when compared to just using individual features extracted. this same combination aim is to demonstrate best unequal treatment among malignant and benign instances. a work concluded a certain incorporating multiple feature extraction methods increase the organizational efficacy after all prostate cancer classification schemes as well as does provide a much more sound foundation regarding computer controlled treatment.

4. m o. shri source text but instead u n. prakash app paired (2017): screening mammography utilizing electronic breast exams resulting outcome data augmentation to attain correct identification. the target of all this survey is really to can use dct - based (dct) coefficients regarding extraction of features but instead implement neurons regarding categorisation. a method of analysis tends to involve trying to transform mammograms photographs into spectral domain employing manual tranny, that encapsulates standout amongst the best

information, accompanied whilst also designation and use a neural networks. one such reach diminishes capture the complexity whereas maintaining critical info. the outcomes increasingly prevalent classification results to greater precision throughout able to distinguish realistic as well as anomalous mucosal. the strategy clearly illustrates efficiency through associated with the management digital innovation. this same authors conclude and it dct-based extracting features blended of neural network models offers an efficient strategy regarding tumor detection as well as elevates diagnosing efficiency.

5. ahmed abdel-nasser alors que abou. (2016): this same precision after all mri images categorisation has been greatly influenced besides visual quality, postprocessing methods, and have regularization. the target of such a study aims to analyse this same impact forward texture-based effectiveness of this approach. this same research methods includes trying different sensor agreements, data pre - processing processes, as well as generalization criteria to test there own impact through classifier performance. the outcomes imply the said decent postprocessing but instead regularization drastically enhance function value but instead categorization results. greater resolution photos provide further detailed info, achieving better. a authors conclude a certain optimise postprocessing actions is important such as increasing the effectiveness like cancer detection structures but also guaranteeing accurate screening results

6. ministry of health (2015):

breast disease remains the leading cause after all mortality globally, trying to highlight need for efficient assessment but instead early sensor arrays. the target of all this document would be to today's global health research but instead underscore the importance like going to address leading causes yeah death, along with disease. its method of analysis tends to involve evaluating global health data from a wide range regions versus recognise developments but instead main risk factors related diseases. its results indicate and it disease accounts for a substantial share of global lives lost, of cancer being among the most typical by many ladies. the outcomes illustrate its vital in need of indicate significant technics but instead advance identification ways to decrease child mortality. a report confirms and it advances in medical technology development, which include automated tracking processes but instead ai - powered techniques, could play a key role through working to improve correct detection but also lowering the worldwide hardship like ovarian cancer.

III. EXISTING SYSTEM

The existing scheme regarding cancer detection relies on such a combined effect like profound feature based to use the vgg19 design but instead categorisation that used a random forest algorithm. calcium (ca2 in just this reach, calcium (ca2 ultrasound of a udders are the first obtained and arranged in to the various categories including in consequential, plus 1 malevolent, plus but also realistic. Calcium (ca2 every photograph has been made smaller to the a classic size (typically 224×224 pixels) versus game this same insert prerequisites of vgg19 internet backbone. Calcium (ca2 this same pre-trained vgg19 design, plus 1 which would be provided with training on even a big set of data, plus 1 is being used as a dimension reduction besides working to remove the latter's connected designation single layer but instead preserving just the convnet core. Plus 1 this permits this same design versus seize elevated temporal but also texture first from images instead of requires significant instruction because after scuff. Plus as once information is extracted, plus 1 they may be

squashed in to the features extracted but instead managed to pass to the a tree based classification. Plus this same classification tree is indeed a training algorithm a certain develops of one hierarchy like rule base based on the feature value systems to categorise a given input in to other various groups. Plus this separates a data - set iteratively premised on some of the most relevant features, plus 1 making it so easy of between perceive as well as incorporate. Plus a classifier is generated utilizing dataset contains but then checked to either invisible sample were collected to judge it's own effectiveness. Plus even though this technique yields a simple but also easy - to - interpret pipe, calcium (ca2 these have damn sure drawbacks like clustering, calcium (ca2 responsiveness of between sounds, calcium (ca2 as well as reduced gross generalisation capacity in comparison with ensemble learning. Calcium (ca2 in consequence, calcium (ca2 while for vgg19 + tree based method could indeed obtain acceptable precision, plus and it might not conduct and also more sophisticated ones and it integrate ensemble methods and maximized classification.

A. Disadvantages of Existing System

The decision forest prototype has been liable to generalization, notably because once given training through strong extracting features that once vgg19, which could also lessen generalized statement to either data sets. It seems to be susceptible versus small differences along data, meaning small variations with in set of data could really bring about a very new tree building but instead inconclusive findings. Compared of between developed ensemble techniques, the choice plant must have fewer parameters but instead stabilisation, starting to make it a little less dependable regarding considerations associated such as medical diagnosis.

IV. PROPOSED SYSTEM

the suggested scheme regarding tumor detection is predicated on even a method that combines deep feature based using vgg19 design with only an supervised learning particular method, plus the additional plants classifiers (etc),plus 1 to realize high precision as well as structural rigidity. Calcium (ca2 in just this structure, plus 1 mammographic photos are all first obtained and arranged in to other prior probability: plus 1 inconsequential, calcium (ca2 malevolent, plus but instead regular. Plus 1 one per photograph seems to be automatically resizes of between 224×224 dots complete fit its feedback specifications of vgg19 design. Calcium (ca2 unless edge detection goggles can be found, plus 1 some kind image compression methodology is utilized besides going to perform element-wise factorial between actual image as well as its correlating disguise, calcium (ca2 whom the enables highlight its cancer area whereas lowering ambient sound. Plus the said helps in improving the quality like input feature as well as means allowing a framework versus specialize in important clinical includes.

the precompiled but instead merged pictures then are did pass thru the pre-trained vgg19 prototype,calcium (ca2 where its final fully connected seem to be deleted,calcium (ca2 but only the convolution layer core was being used for extracting features.plus 1 the said stage evolves one per photograph it in to a top feature the said encapsulates powerful image character traits including such surface,plus 1 back edge,plus as well as functional formations related to different types after all dense breast.calcium (ca2 those same extracted the features seem to be saved together with about there correct class stickers and seem to be additional divided up complete delete whatever gon na order leanings.plus 1 its set of data then is separated

in to other training components,calcium (ca2 generally to use an 125:approximately 25 proportion,calcium (ca2 guaranteeing that now the design seems to be reviewed through data sets.

A. System Architecture

The proposed system would be constructed as both a organized as well as customizable pipe the said comes ieee standard rules, making sure clear communication, cloud computing, as well as efficient information stream among elements. this same design usually starts with input data, at which mammographic photos alongside there own relating edge detection headgear were also procured and arranged in and out of predetermined classrooms.

B. Preprocessing Pipeline

the image preprocessing component seems to be a vital stage a certain morphs uncooked ultrasound in to the impactful image features useful regarding categorisation.plus 1 orginally,plus 1 every one of image pixels as well as about there roughly equivalent classification hoods seem to be gathered that once class-specific directories,plus order to ensure just legitimate material ” seem to be digested but exclude helmet documents from of the principal photo identify.calcium (ca2 so every photo would be automatically resizes to either a basic resolving yeah 224×224 pixel resolution versus game its feedback prerequisites of vgg19 deep convolutional neural network.plus its correlating conceal,calcium (ca2 equipped through gray - scale,calcium (ca2 seems to be mainstreamed to the a range from 0 versus but then transferred into such a three-channel template complete coincide the with color space framework of image pixels.plus 1 a crucial move through data pre - processing would be single image,calcium (ca2 in which element-wise factorial seems to be managed to perform between both the actual image as well as its helmet,plus efficaciously emphasizing a cancer province but also attempting to suppress inconsequential basic information.

C. Software & Hardware Requirements

Software: Windows 11, Python, TensorFlow 2.x, Keras, NumPy, Pandas, OpenCV, Scikit-learn, Matplotlib. Hardware: Intel Core i5 / Pentium IV 2.4 GHz processor, 8 GB RAM (minimum), NVIDIA GPU (recommended), 40 GB Hard Disk storage.

D. Advantages of Proposed System

provides high precision but also generalisation besides blending deep convolutional excavation to transfer learning. reduces performance of the model thanks to randomisation along features extraction but instead number of decision plants. handles tall data effectively, making it ideal such as complicated photo categorization tasks.

V. RESULTS AND DISCUSSIONS

the results but also discourse demonstrates that the suggested framework utilising vgg19 as for additional amount tree branches classification model shows better performance when compared with the present designs. its model ensures high precision alongside centered exactness, actually remember, but also score, whom the signifies it can clearly identify all malignant but instead quasi instances. its original dataset demonstrates too few discrepancies in all classrooms, and indeed the curve demonstrated excellent school disconnection capacity. in comparison with classification tree as well as reinforcement learning, its model building accomplishes so much constantly as well as controls various image characteristics extra successfully because of it's own outfit environment. its use of image compression both helps improve recognition whilst also specializing in essential cancer cell territories. average, the outcomes affirm that now the structure would be trustable, efficient, but instead suitable regarding continuing to support breast cancer diagnosis.

A. Classification Performance

The decision tree–based classification framework demonstrates strong and reliable performance across ovarian cancer subgroups, achieving an overall accuracy of about 95.16% with consistently high precision (95.11%) and recall (95.07%), indicating that the model is both effective in correctly identifying positive cases and dependable in minimizing missed detections. The balanced F1-score further reflects a stable trade-off between precision and recall, confirming robustness in classification behavior. Confusion matrix analysis shows that most normal and malignant samples are correctly classified, with only a small number of misclassifications occurring between benign and malignant categories, which suggests that the model occasionally struggles with distinguishing visually or feature-wise similar classes. The ROC analysis supports this outcome by showing a curve close to the top-left region with an AUC of approximately 96.01%, indicating strong separability between classes and low false-positive rates overall. Similarly, the Extra Trees Classifier also exhibits competitive performance across all three classes, with high classification reliability for normal and malignant cases, while showing minor confusion in borderline benign cases

Parameter	Decision Tree (DT)	Extra Trees Classifier (ETC)
Classification Accuracy	94.14%	98.10%
Precision	95.16%	98.00%
Recall / Sensitivity	95.07%	98.00%
F1-Score	95.11%	98.04%
ROC-AUC Score	96.01%	98.48%

B. Comparative Analysis

The classification model demonstrates excellent performance across ovarian cancer stages, achieving an overall accuracy of 98.1% with consistently high precision (98.0%), recall (98.0%), and F1-score (98.04%), indicating a well-balanced and reliable predictive system. Class-wise results further highlight strong discrimination capability, with benign (normal) cases achieving 97.8% accuracy and malignant cases reaching 98.4%, showing that the model performs slightly better in identifying malignant instances while maintaining very low misclassification rates for both categories.

Class / Stage	Accuracy	Precision	Recall	F1-Score
Benign (Normal)	97.8%	98.0%	97.5 %	97.7 %
Malignant	98.4%	98.2%	98.6 %	98.4 %
Overall Model	98.1%	98.0%	98.0 %	98.04 %

C. Test Cases

Table IV presents the functional test cases used to validate the proposed breast cancer detection system under various operational scenarios.

TABLE IV. System Test Cases

S.No	Input	If Available	If Not Available
1	Upload breast cancer dataset	Dataset loaded successfully	No process
2	Data preprocessing (cleaning & normalization)	Processed dataset generated	No process
3	Generate train & test split	Training and testing sets created	No process
4	Run ML/DL classification model	Model trained and results displayed	No process
5	Perform feature selection/analysis	Important features identified	No process

6	Accuracy and performance evaluation	Metrics (accuracy, precision, recall, F1-score) displayed	No process
7	Confusion matrix generation	Classification matrix displayed	No process

VI. CONCLUSION

Using a combination of deep learning and ensemble machine learning methods, the study presents a method for the accurate and dependable detection of breast cancer. The system successfully extracts complex patterns and texture information from ultrasound images by utilizing the VGG19 model for feature extraction. The quality of the extracted features is enhanced by the integration of image fusion, which further intensifies the focus on tumor regions. Due to its ensemble nature, which reduces overfitting and enhances generalization, the Extra Trees Classifier performs better than any other model in the evaluation.

Reference

1. Mohamed Abdel-Nasser, Antonio Moreno, Mohamed A. Abdel wahab, Adel Saleh, Saddam Abdulwahab 1, Vivek K. Singh and Domenech Puig, “Matching Tumour Candidate Points in Multiple Mammographic Views for Breast Cancer Detection ”, *2019 International Conference on Innovative Trends in Computer Engineering (ITCE'2019)*, Aswan, Egypt, 2019.

2. F. F. Ting, K. S. Sim, “Self-regulated Multilayer Perceptron Neural Network for Breast Cancer Classification ”, *International Conference on Robotics, Automotion and Sciences (ICORAS)*, 2017.

3. Mohamed A. Berbar, “Hybrid methods for feature extraction for breast masses classification ”, *Egyptian Informatics Journal*, 2017.

4. Assistant Prof T. Krishna Chaitanya, P. Chandra Sekhar Azad, “Neural Network Based Classification of Digital Mammograms using DCT Coefficients ”, *International Journal of Advance Engineering and Research Development*, 2017.

5. Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, and Domenec Puig, “The Impact of Pixel Resolution, Integration Scale, Pre-processing, and Feature Normalization on Texture Analysis for Mass Classification in Mammograms ”, *Hindawi Publishing Corporation International Journal of Optics Volume* 2016.

6. *The top 10 causes of death. World Health Organization*. http://www.who.int/mediacentre/factsheets/fs310/en/. Accessed, 2015

Author Details.

Ms.M.Priyanka completed her Master of Computer Applications (MCA) from Acharya Nagarjuna University. She is currently working as an Assistant Professor. Her teaching subjects include C Programming, Data Structures, Java, and Computer Networks. Her areas of interest include Programming, Machine Learning, and Networking

Ms.K.Pavani Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with MCA

,M. Tech in Computer Science .She has 10 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada,NTR District. Her area of interest includes AI ML, etc

Ms.Sk.Nasrin is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc.(Mathematics, Statistics ,computer science) from SDM Siddhartha Mahila Kalasala. Her area of interest are DBMS and Machine Learning with Python.