Breast Cancer Detection Using Machine Learning Techniques

[1]Anju Kumari, [2]Mohana Vani S, [3]Namitha V Pawar, [4]Vijay Kumar S,[4]

Assistant Professor
[1]Information Science and Engineering Department,
[1]BNM Institute of Technology, Bangalore, India

Abstract— Breast cancer is the second leading cause of demise for women, so correct early detection can assist lower breast most cancers mortality prices. This paper targets at giving an overview of breast most cancers detection using machine learning techniques. The methodology used here is a Convolutional Neural network with a pre-trained model. For the implementation of the ML algorithms, the dataset was partitioned into the schooling section and the checking out section.

***Index Terms*—System Design and Development, Machine Learning, Feature Extraction.**

I. INTRODUCTION

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes. Machine learning algorithms use historical data as input to predict new output values .The Internet of things describes physical objects that are embedded with sensors. It enables various applications such as Breast Cancer Detection Using Machine Learning Techniques. Cancer is a creation of abnormal cells that come from a modification in these cells genetically and spreads into the body, a late in diagnosis and treatment leads to death. There are two types of breast cancer, invasive and non-invasive. The former is harmful, malignant, ability to infect other organs, and classified as cancerous.

II. LITERATURESURVEY

Breast Cancer can be detected using various methods and algorithms. In the paper [1], Deep Neural Network with SupportValue (DNNS) is introduced to produce better quality images and to fix other performance parameters. The authors propose a newalgorithm or pseudocode along with mathematical formulas to evaluate the efficiency and performance. Unlike other methods, theproposed method is based on Support value on a deep neural network. To meet the better performance, efficiency, and quality ofimages, a normalization process has been employed. Experimental results proved that the proposed DNNS is quite better than theexisting methods. It is ensured that the proposed algorithm is advantageous in both performance, efficiency and quality of imagesare crucial in the latest medical systems. Deep Neural networks usually require much more data than traditional machine learningalgorithms. Deep neural networks are also more computationally expensive than traditional algorithms in machine learning. Thereis no standard theory to guide you in selecting right deep learning tools as it requires knowledge of topology, training method andother parameters. In this paper [2], Proposed methodology in paper is ML methods used were: Artificial Neural Network (ANN),ExtremeLearningMachine(ELM),SupportVectorMachine(SVM)andK-NearestNeighbors(k-NN)Inaddition,thehyperparameter values giving the least errors for ANN, ELM, k-NN and SVM methods are determined using Hyperparameteroptimization technique. The importance of this work is pretty high because of the usage of the different type of data. In addition,this study is also important because four different ML methods are compared. In addition, this study may support the further workin this field. The obtained accuracy rate cannot be regarded as very high. The k-NN method does not actually contain the trainingphase. [3], Proposed methodology in this paper is seven phases of Machine Learning are used in this paper. They are: Pre-Processing Data, Data preparation, Features Selection, Feature Projection, Feature Scaling, Model Selection and Prediction.Methods used are: Logistic Regression, k Nearest Neighbor (k-NN), Support Vector machine, thehighestvalueofcorrectlyclassifiedinstancesandthelowervalueofincorrectlyclassifiedinstancesthantheotherclassifiers.K-NNis a lazy learner and does not do much during training process unlike others classifiers that build the models. Future enhancementsin this paper is that further research in this field should be carried out for the better performance of the classification techniques sothat it can predict on more variables. It is needed to reduce the error rates with maximum accuracy. [4], Proposed methodology inpaper this paper is Data Mining and Machine Learning Algorithms. They are: Decision tree algorithms, K-nearest-neighbours(kNN) algorithm, Support Vector Machine (SVM), Naïve Bayes (NB) It is a probabilistic classifier and Logistic regression. UCIMachine Learning Repository for breast cancer dataset is used in this project. It is observed that a good dataset provides betteraccuracy. Data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extractingknowledge from such suitable databases. K-NN is a lazy learner and does not do much during training process unlike othersclassifiers that build the models. Future enhancement of this paper is Selection of appropriate algorithms with good home datasetwill lead to the development of prediction systems. [5], Proposed methodology in paper is the Search strategy, Inclusion and exclusion criteria, Data extraction, prediction model risk of bias assessment tool(PROBAST)this is the first system aticre view of the application of ML to breast cancer survival prediction, and accurate 5-year survival predictions are very important for furtherresearch. The models still face limitations related to a lack of data preprocessing steps, the excessive differences of sample featureselection, and issues related to validation and promotion. Future enhancement of this paper is the modelperformancestillneedstobefurtheroptimized,ndotherbarriersshouldbeaddressed.Researchersandmedicalworkersshouldconnectwithreality,choosea model carefully, use the model in clinical practice after verification, and use rigorous design and validation methods with a largesampleofhigh-qualityresearchdataonthebasisofpreviousfindings.

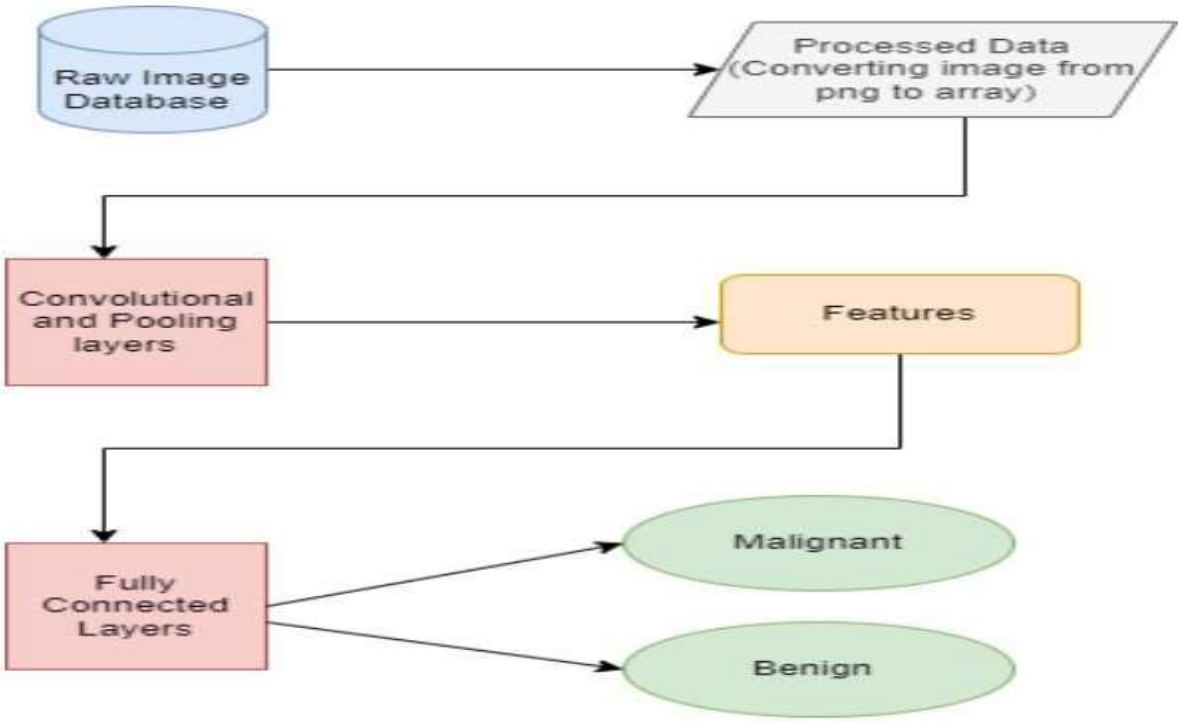
III. DATASETUSED

The dataset which is used in this model is the Breast Cancer Histopathological Database (BreakHis) dataset. The Breast Cancer

HistopathologicalImageClassification (BreakHis) iscomposedof9, 109microscopicimagesofbreasttumortissuecollectedfrom82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). This database has been built incollaborationwiththeP&DLaboratory –PathologicalAnatomyandCytopathology,Parana,Brazil.

IV. SYSTEMDESIGNANDDEVELOPMENT

Architectural Design: Requirements of the software should be transformed into an architectural that describes the software'stop-level structure and identifies its components. This is accomplished through architectural design (also called system design), whichactsasapreliminaryblueprintfromwhichsoftwarecanbedeveloped.IEEEarchitecturaldesignastheprocessofdefiningacollection of hardware and software components and their interface to establish the framework for the development of a computersystem. This framework is established by examining the software requirement document and designing a model for providingimplementation details. These details are used to specify the components of the system along with their inputs, outputs, functions, andtheinteractionbetween them.Anarchitecturaldesignperformsseveralfunctions.



ArchitecturalDesignoftheModel

V. IMPLEMENTATION

ListofModulesTheprojecthasbeenimplementedbydividingtheentireprojectintothreemodules.Theyare:

- Training
- Utilities
- Testing

➢ Training: The training file follows an abstraction model of programming. This file systematically calls all the functionsrequiredfortheprojecttoexecutesuccessfullyandcreateamodel.Followingamodularapproach,thisfilehasmanystepso rfunctioncallsaslistedbelow:

- Importingdata
- Datavisualizationanddistribution
- Preparingdataforprocessing
- Splittingofdata
- Addingmorevarietyandvariancetothedata
- Preprocessing
- Creatingabatchgeneratorfortraining
- Creatingthemodel
- Trainingthemodel
- Savingmodelondisk
- Modelanalytics

Algorithm:Thissectionexplainsindetailabouttheconceptofconvolutionalneuralnetworksandlatertheproposedneuralnetworkmodel usedintheprojectandthatistalkedaboutinthesection,alongwithimportantlibrariesusedthroughouttheproject.

Convolutional Neural Network CNN: comprises of one or more convolutional layers. It is proceeded by one or more neuralnetwork connected layers. The purpose of this layer is to receive a feature map. Usually, we start with low number of filters forlow-levelfeaturedetection.ThedeeperwegointotheCNN,themorefiltersweusetodetecthigh-levelfeatures.Featuredetectionis based on 'scanning' the input with the filter of a given size and applying matrix computations in order to derive a feature

map.Poolingisamethodofarbitraryexperimentinganditiscommonlyusedwhileaddingpoolinglayerstolowertheparametersandtoeliminateunnecessaryfeaturesduringthetraininginapursuancetoavoidoverfittinginthenetwork.Aftertheconvolutionallayers,themultiplepropportionalarraysareflattenedintoatwo-dimensionalarrayinfullyconnectednetwork.

Pooling Layer: Similar to the Convolutional layer, the pooling layer is responsible for reducing the spatial Size of theconvolved feature. This is to decrease the computational power required to process the data through dimensionality reduction.Furthermore,tisusefulforextractingdominantfeatureswhicharerotationalandpositionalinvariant,thusmaintainingtheprocessofeffectivetrainingofthemodel.

Fully Connected Layer: In a fully connected layer, we flatten the output of the last convolution layer and connect every nodeofthecurrentlayerwiththeothernodesofthenextlayer.Neuronsinafullyconnectedlayerhavefullconnectionstoallactivationsinthepreviouslayer,asseeninregularNeuralNetworksandworkinasimilarway.Addingafullyconnectedlayerisacheapwayof learning non-linear combinations of the high-level features which can be represented by the output of the convolutional layer.Thefullyconnectedlayerislearningapossiblynon-linearbyflatteningtheimageintoacolumnvector.

TensorFlow: TensorFlowisalicensefreesoftwarefordataflowgraphstobuildmodels.Itallowsneuralnetworkstogohandinhand with multi layers and is predominantly used for prediction and classification of the sample data fed into the model or thenetwork. TensorFlow provides excellent functionalities and services and enables the high level complex parallel computation forbuildingadvancedneuralnetworks.

Keras: It is also another python library majorly used in neural networks. It is more user friendly, modular and enables fasterexperimentation with neural networks. It is can be made to run on top of Tensor flow. It contains a great number of neural networkbuildingblockssuchasactivationfunctions, optimizers, andtoolscompatiblewithworkingoftextandimagedata.

Flask: It is a micro web framework written in Python. It is classified as a microframework because it does not requireparticulartoolsorlibraries.Ithasnodatabaseabstractionlayer,formvalidation,oranyothercomponentswherepre-existingthird-party libraries provide common functions. However, Flask supports extensions that can add application features as if they wereimplemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various openauthenticationtechnologiesand severalcommonframeworkrelatedtools.

- **Utilities**

  This python file contains the main logic of the entire project. Every task is executed in a different function, making the codemodularandalsoattainingagoodlevelofabstraction.Allthefunctioncallsaremadeinthetrainingfileoftheproject,asseeninsection

- **Testing**

Thispythonscriptusesthealreadycreatedandsavedmodel.Aserversocketiscreatedwhichlistenstotheappropriateport.Thecode in this file is responsible for taking in the pictures from the simulator, preprocess it, pass it through the neural network'smodelandsendthegeneratedaccelerationandsteeringanglevaluesbacktothesimulator'scaragent.

VI. CONCLUSIONANDSCOPEFORFUTUREENHANCEMENT

The diagnosis of breast cancer in an early stage can help in the reduction of the mortality caused by breast cancer. In thisproject, we have demonstrated how to classify benign and malignant breast cancer from a collection of microscopic images usingconvolutional neural networks. We implemented pre-trained CNN models with fine tuning leveraging transfer learning to observetheclassificationperformanceofbreastcancerfrommicroscopicimages.Weevaluatedthefine-tunedpre-trainedmodelsapplyingResNet50 with the Adam optimizers. Although this project is far from complete but it is remarkable to see the success of deeplearning in such varied real world problems. We have also proposed several strategies for training the CNN architecture, based ontheextractionofpatchesobtainedrandomlyorbyaslidingwindowmechanism, thatallowtodealwiththehigh-resolutionofthesetextured images without changing the CNN architecture designed for low-resolution images. Our experimental resultsobtainedontheBreaKHisdatasetshowedimprovedaccuracyobtainedbyCNNwhencomparedtotraditionalmachinelearningmodel strainedonthesamedatasetbutwithstateofthearttexturedescriptors.

Future work can explore different CNN architectures and the optimization of the hyperparameters. Also, strategies to selectrepresentative patches in order to improve the accuracy can be explored and also to validate the model with other datasets thatincludenew images.

REFERENCES

[1] P.BoyleandB.Levin,Eds.,WorldCancerReport2008.Lyon:IARC,2008.[Online].Available:http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdfG.Eason,B.Noble,andI.N.Sneddon,"Oncertain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol.A247, pp.529–551,April 1955.*(references)*

[2] S. R. Lakhani, E. I.O., S. Schnitt, P. Tan, and M. van de Vijver, WHO classification of tumours of the breast, 4th ed.Lyon:WHOPress, 2012S.

[3] B. Stenkvist, S. Westman-Naeser, J. Holmquist, B. Nordin, E. Bengts son, J. Vegelius, O. Eriksson, and C. H. Fox,"Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," CancerResearch,vol. 38, no. 12, pp.4688–4697,1978.

[4] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer basedonfineneedlebiopsy microscopicimages,"Computersin Biology and Medicine,vol.43,no.10,pp.1563–1572,2013.

[5] P. Filipczuk, T. Fevens, A. Krzyz´ak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis ofcytologicalimagesoffineneedlebiopsies,"IEEETransactionsonMedicalImaging,vol.32,no.12,pp.2169–2178,2013.

[6] Y. M. George, H. L. Zayed, M. I. Roushdy, and B. M. Elbagoury, "Remote computer-aided breast cancer detection anddiagnosissystembasedon cytological images," IEEE SystemsJournal, vol. 8,no.3, pp. 949–964,2014.

[7] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable randomsubspaceclassifier ensembles," MachineVisionandApplications, vol.24, no.7, pp.1405–1420,2013.

[8] Y.Zhang,B.Zhang,F.Coenen,J.Xiau,andW.Lu,"One-classkernelsubspaceensembleformedicalimageclassification,"EURASIP Journalon Advancesin SignalProcessing,vol. 2014,no. 17,pp. 1–13,2014.

[9] S.Doyle,S.Agner,A.Madabhushi,M.Feldman,andJ.Tomaszewski,"Automatedgradingofbreastcancerhistopathology using spectral clustering with textural and architectural image features," in Proceedings of the 5th IEEEInternationalSymposiumonBiomedicalImageging (ISBI):From Nano to Macro,vol.61.IEEE,May2008,pp.496–499.

[10] A. J. Evans, E. A. Krupinski, R. S. Weinstein, and L. Pantanowitz, "2014 american telemedicine association clinicalguidelines for telepathology: Another important step in support of increased adoption of telepathology for patient care,"JournalofPathologyInformatics, vol. 6,2015.