WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detection of Breast Cancer Based on Machine Learning Using Ensemble of Classifiers

Mrs. Gowri Subadra K[1], Harshika D[2], Purvashree A[3], Shaveetha R[4]

Assistant Profesor[1], Student[2,3,4]

[1,2,3,4]Sri Manakula Vinayagar Engineering College, Puducherry, India.

ABSTRACT-

Breast Cancer is one of the leading causes of death among women. Early detection of the Breast cancer and effective treatment is crucial for the improvement of survival rate. The effects of cancer can be reduced if detected early. Many methods were found to detect the Breast cancer however these existing methods still need improvement. We proposed a system enabling the healthcare industry to detect breast cancer quickly and accurately. Machine learning is widely used in breast cancer pattern classification due to its advantages in modelling a critical feature detection from complex Breast cancer datasets. A system for automatic detection of Breast cancer diagnosis and prognosis using ensemble of classifiers. We review various machine learning algorithms and ensemble of different ML algorithms. We present an ensemble of different classifiers for automatic Breast cancer diagnosis and prognosis detection. We also present and compare various ensemble models and other variants of tested Machine learning based models with and without up-sampling technique on two benchmark datasets. The results showed that the ensemble methodology outperformed different progressive strategies and achieved 98.8% accuracy. It offers high performance

KEYWORDS- Diagnosis and Prognosis detection, Benchmark datasets.

I. INTRODUCTION

Breast cancer could be a style of cancer that begins within the breast tissue. It happens once cells within the breast begin to grow and divide in an uncontrolled approach, eventually forming a neoplasm. If left untreated, willcer carcinoma} cells can unfold to alternative components of the body, as well as the body fluid nodes, bones, and organs. There are many sorts of carcinoma, as well as ductal cancer, lobe cancer, and inflammatory carcinoma. The foremost common kind is ductal cancer that starts within the cells that line the milk ducts within the breast. Lobe cancer starts within the cells that turn out milk, and inflammatory carcinoma could be a rare style of carcinoma that affects the skin and body fluid vessels within the breast. Breast cancer can occur in each men and women, however it's rather more common in girls. Risk factors for carcinoma embrace being feminine, being over the age of fifty, having a case history of carcinoma, being overweight or weighty, and having a private history of carcinoma or alternative sorts of cancer. There are many treatment choices for carcinoma, as well as surgery, therapy, action therapy, and secretion medical aid. The particular treatment arrange can rely on the kind and stage of the cancer, further more because the patient's overall health and preferences. Early detection and treatment of willcer carcinoma can considerably improve the probabilities of a in outcome. It's vital for people to remember of their own bodies and to report any changes or abnormalities to a tending supplier. This will embrace activity self-exams and obtaining regular mammograms, furthermore as following suggested pointers for carcinoma screening

II. MACHINE LEARNING ALGORITHM

Machine learning is a field of artificial intelligence that focuses on the development of algorithms and models that can learn from and make predictions or decisions based on data. It involves training a computer program using a large dataset, allowing the program to improve its performance at a given task by adjusting the algorithms and models it uses based on the data it has seen. One of the key goals of machine learning is to enable computers to learn and adapt to new situations without explicit programming. This is in contrast to traditional programming, where a developer writes code to

perform a specific task or solve a specific problem. With machine learning, the computer is able to learn from the data it is given and make decisions or predictions on its own.

There are many different types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning. Each of these approaches involves using different algorithms and techniques to train the machine and improve its performance. Machine learning has many practical applications, including in image and speech recognition, natural language processing, and predictive modelling.

There are several different types of machine learning, each of which involves using different techniques and approaches to train a machine to learn and make decisions. These types include:

1. Supervised learning: This involves training a machine on a labelled dataset, where the correct output is provided for each example in the training set. The machine uses this labelled data to learn a function that maps input data to the correct output.
2. Unsupervised learning: This involves training a machine on an unlabelled dataset, where the correct output is not provided. The machine must learn to identify patterns and relationships in the data on its own.
3. Semi-supervised learning: This involves training a machine on a dataset that is partially labelled, with some examples in the training set having the correct output provided and others not. This can be useful when it is expensive or time-consuming to label all of the data.
4. Reinforcement learning: This involves training a machine to make decisions in an environment in order to maximize a reward. The machine learns through trial and error, adjusting its actions based on the rewards it receives.
5. Transfer learning: This involves training a machine on one task and then fine- tuning it for a related task. This can be useful when there is a limited amount of data available for the new task.

a) Support Vector Machine

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. They are based on the idea of finding a hyper plane in a high-dimensional space that maximally separates the different classes.In the case of classification, an SVM algorithm will take a dataset as input and output a classifier that can be used to predict the class of a new data point. The classifier is a line (in two dimensions) or a plane (in three dimensions) that divides the data points into different classes.

SVMs are particularly useful in cases where the number of dimensions (i.e. the number of features) is much greater than the number of samples. They are also effective in cases where the data is highly imbalanced (i.e. there are many more data points in one class than the other). One of the key advantages of SVMs is that they are highly flexible and can be used with a wide range of kernel functions, which allows them to learn complex decision boundaries. They are also relatively robust to over fitting, which makes them a good choice for many real-world applications.

b) Logistic regression

Logistic regression is a type of supervised learning algorithm that is used to predict a binary outcome, such as 0 or 1, true or false, or yes or no. It is based on the concept of logistic function, which maps the input data to a value between 0 and 1. In logistic regression, the goal is to learn a function that takes in a set of features and outputs a probability that the outcome is 1. This probability is then used to predict the class of a new data point. For example, if the probability that the outcome is 1 is greater than 0.5, the model will predict that the outcome is 1, and if the probability is less than 0.5, the model will predict that the outcome is 0.

Logistic regression is a linear model, which means that it makes a prediction based on a linear combination of the input features. It is useful in cases where the relationship between the input features and the output is approximately linear. It is also relatively simple to implement and can be trained efficiently using gradient descent. Logistic regression is often used in a variety of applications, including spam filtering, credit risk assessment, and medical diagnosis.

c) Naive Bayes

Naive Bayes is a type of supervised learning algorithm that is based on the idea of Bayesian probability. It is often used in natural language processing and spam filtering, and is particularly useful for classification tasks with a large number of features. The algorithm works by making predictions based on the probability of an event occurring, given the occurrence of certain features. It makes the assumption that the features are independent of each other (hence the name "naive"), which allows it to make predictions based on the individual probabilities of each feature. One of the key advantages of Naive Bayes is that it is relatively simple and easy to implement, and it can be trained quickly even on large datasets. It is also relatively robust to the presence of irrelevant features, as these are assumed to be independent of the output and therefore do not affect the prediction. However, the assumption of independence can sometimes lead to suboptimal performance in certain cases.

d) Decision tree

Decision trees are a type of machine learning algorithm that are used for both supervised and unsupervised learning. They are called decision trees because they involve creating a tree-like model of decisions, with the branches representing possible decisions and the leaves representing the final outcome. In the case of supervised learning, a decision tree is trained on a labelled dataset, where the correct output is provided for each example in the training set. The tree is created by analysing the features of the training data and determining which feature is the most important for predicting the output.

This process is repeated recursively until the tree is fully grown. Once the decision tree has been trained, it can be used to make predictions about the output for new data points. This is done by traversing the tree and making a decision at each node based on the value of the feature at that node. The final prediction is made at the leaf node. Decision trees are often used in a variety of applications, including credit risk assessment, medical diagnosis, and customer segmentation. They are relatively simple to understand and interpret, and they can handle both numerical and categorical data. However, they can be prone to over fitting if the tree is allowed to grow too deep.

e) Artificial neural network

An artificial neural network (ANN) is a type of machine learning model that is inspired by the structure and function of the brain. It is composed of a large number of interconnected "neurons," which are inspired by the neurons in the brain and are used to process and analyse data. In an ANN, the neurons are organized into layers, with the input layer receiving the raw data and the output layer producing the final predictions or decisions. Between the input and output layers, there are one or more hidden layers, which are used to extract features and patterns from the data. The connections between the neurons are weighted, and the weights are adjusted during the training process to optimize the performance of the model.

Deep learning algorithms are a type of artificial neural network that involve training a model with multiple hidden layers. These hidden layers allow the model to learn complex patterns and features in the data, and they are what gives deep learning algorithms their name. Artificial neural networks have been successful in a variety of tasks, including image and speech recognition, natural language processing, and machine translation. They are able to learn from large amounts of data and make highly accurate predictions or decisions. However, they can be computationally intensive and require large amounts of data to train effective.

III. LITERATURE SURVEY

[1]. Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, Kenny H. Cha, In this work we developed a deep convolutional neural network (CNN) for classification of malignant and benign masses in digital breast tomosynthesis (DBT) using a multistage transfer learning approach that utilized data from similar auxiliary domains for intermediate-stage fine- tuning. Breast imaging data from DBT and digitized screen-film mammography (SFM), digital mammography (DM) totaling 4,039 unique ROIs (1,797 malignant and 2,242 benign) were collected. Using crossvalidation, we selected the best transfer network from six transfer networks by varying the level up to which the convolutional layers were frozen. In a single-stage transfer learning approach, knowledge from CNN trained on ImageNet data was fine-tuned directly with DBT data. In a multi-stage transfer learning approach, knowledge learned from ImageNet was first fine-tuned with the mammography data and then fine-tuned with the DBT data. Two transfer networks were compared for the secondstage transfer learning by freezing most of the CNN structure versus freezing only the first convolutional layer. We studied the dependence of the classification performance on training sample size for various transfer learning and fine-tuning schemes by varying the training data from 1% to 100% of the available sets. The area under the receiver operating characteristic curve (AUC) was used as a performance measure. The view-based AUC on the test set for single-stage transfer learning was 0.85±0.05 and improved significantly.

[2]. Bo Fu, Pei Liu, Jie Lin, Ling Deng, Kejia Hu, Hong Zheng, Chinese women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare an appropriate treatment plan that may prolong patient survival time. An alternative prognosis model framework to predict Invasive Disease-Free Survival (iDFS) for early-stage breast cancer patients, called MP4Ei, is proposed. MP4Ei framework gives an excellent performance to predict the relapse or metastasis breast cancer of Chinese patients in 5 years. Methods: MP4Ei is built based on statistical theory and gradient boosting decision tree framework. 5246 patients, derived from the Clinical Research Center for Breast (CRCB) in West China Hospital of Sichuan University, with early-stage (stage I-III) breast cancer are eligible for inclusion. Stratified feature selection, including statistical and ensemble methods, is adopted to select 23 out of the 89 patient features about the patient' demographics, diagnosis, pathology and therapy. Then 23 selected features as the input variables are imported into the XGBoost algorithm, with Bayesian parameter tuning and cross validation, to find out the optimum simplified model for 5-year iDFS prediction. Results: For eligible data, with 4196 patients (80%) for training, and with 1050 patients (20%) for testing, MP4Ei achieves comparable accuracy with AUC 0.8451, which has a significant advantage ($p < 0.05$).

[3]. ZHIQIONG WANG, MO LI, HUAXIA WANG, HANYU JIANG, YUDONG YAO (Fellow, IEEE), HAO ZHANG, AND JUNCHANG XIN, A computer-aided diagnosis (CAD) system based on mammograms enables early breast cancer detection, diagnosis, and treatment. However, the accuracy of existing CAD systems remains unsatisfactory. This paper explores a breast CAD method based on feature fusion with Convolutional Neural Network (CNN) deep features. First, we propose a mass detection method based on CNN deep features and Unsupervised Extreme Learning Machine (US- ELM) clustering. Second, we build a feature set fusing deep features, morphological features, texture features, and density features. Third, an ELM classifier is developed using the fused feature set to classify benign and malignant breast masses. Extensive experiments demonstrate the accuracy and efficiency of our proposed mass detection and breast cancer classification method.

[4]. XINGYU LI 1, (Member, IEEE), MARKO RADULOVIC2, KSENIJA KANJER2, AND KONSTANTINOS N. PLATANIOTIS1, (Fellow, IEEE), Accurate diagnosis of breast cancer in histopathology images is challenging due to the heterogeneity of cancer cell growth as well as a variety of benign breast tissue proliferative lesions. In this paper, we propose a practical and self-interpretable invasive cancer diagnosis solution. With minimum annotation information, the proposed method mines contrast patterns between normal and malignant images in a weak-supervised manner and generate a probability map of abnormalities to verify its reasoning. Particularly, a fully convolutional autoencoder is used to learn the dominant structural patterns among normal image patches. Patches that do not share the characteristics of this normal population are detected and analyzed by one-class support vector machine and one-layer neural network. We apply the proposed method to a public breast cancer image set. Our results, in consultation with a senior pathologist, demonstrate that the proposed method outperforms existing methods. The obtained probability map could benefit the pathology practice by providing visualized verification data and potentially leads to a better understanding of datadriven diagnosis solutions.

[5]. Morteza Heidari, Sivaramakrishnan Lakshmivarahan, Seyedehnafiseh Mirniaharikandehei, Gopichandh Danala, Sai Kiran R. Maryada, Hong Liu, Bin Zheng, Since computer-aided diagnosis (CAD) schemes of medical images usually computes large number of image features, which creates a challenge of how to identify a small and optimal feature vector to build robust machine learning models, the objective of this study is to investigate feasibility of applying a random projection algorithm (RPA) to build an optimal feature vector from the initially CAD-generated large feature pool and improve performance of machine learning model. Methods: We assemble a retrospective dataset involving 1,487 cases of mammograms in which 644 cases have confirmed malignant mass lesions and 843 have benign lesions. A CAD scheme is first applied to segment mass regions and initially compute 181 features. Then, support vector machine (SVM) models embedded with several feature dimensionality reduction methods are built to predict likelihood of lesions being malignant. All SVM models are trained and tested using a leave-one-case-out cross-validation method. SVM generates a likelihood score of each segmented mass region depicting on one-view mammogram. By fusion of two scores of the same mass depicting on two-view mammograms, a case-based likelihood score is also evaluated. Results: Comparing with the principle component analyses, nonnegative matrix factorization, and Chi-squared methods, SVM embedded with RPA yielded a significantly higher casebased lesion classification performance with the area under ROC curve of 0.84±0.01 ($p<0.02$).

IV. CONCLUSION

We proposed a method for breast cancer detection using machine learning techniques. Benchmark datasets are used for the experiments. Classifiers based on machine learning and deep learning to increase classification and prediction accuracy. Several ensembles of different ML-based classifiers were also tested for the classification of BC. We found out that SVM outperforms both datasets compared to all ML classifiers and ANN from DL classifiers when used individually. For the ensembling method, performs well without and with up sampling on the diagnosis dataset, whereas outperforms all other combinations on the prognosis dataset when ANN is used as a final layer. We also observed an increase in performance when balanced class weights are used along with the up sampling technique as compared to without, and the up sampling technique is used individually. The performance was also analysed using a different number of K-fold for the best ensemble classifier. In the future, we intend to apply more advanced models for the automatic detection of BC.

REFERENCE

[1]. IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. X, DECEMBER 2017. Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, Kenny H. Cha

[2]. IEEE Transactions on Biomedical Engineering, DOI 10.1109/TBME.2018.2882867, TBME-01194- 2018.R1, Bo Fu*, Pei Liu, Jie Lin, Ling Deng, Kejia Hu, Hong Zheng*

[3]. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2017.DOI: 10.1109, ZHIQIONG WANG1, 5, MO LI2, HUAXIA WANG3, HANYU JIANG3, YUDONG YAO3 (Fellow, IEEE), HAO ZHANG4, AND JUNCHANG XIN2.

[4]. Received February 20, 2019, accepted March 7, 2019, date of publication March 11, 2019, date of current version April 2, 2019. *Digital Object Identifier 10.1109/ACCESS.2019.2904245* XINGYU LI 1, (Member, IEEE), MARKO RADULOVIC2, KSENIJA KANJER2, AND KONSTANTINOS N. PLATANIOTIS1, (Fellow, IEEE)

[5]. IEEE Transactions on Biomedical Engineering, DOI 10.1109/TBME.2021.3054248, TBME-02405- 2020, Morteza Heidari, Sivaramakrishnan Lakshmivarahan, Seyedehnafiseh Mirniaharikandehei, Gopichandh Danala, Sai Kiran R. Maryada, Hong Liu, Bin Zheng

[6]. IEEE Xplore Part Number: CFP21OAB- ART; ISBN: 978-1-7281-9537-7, Yogesh Suresh Deshmukh, Parmalik Kumar, Rajneesh Karan, Sandeep K. Singh

[7]. IEEE 49239, a Comprehensive Study of Machine Learning Approach on Cytological Data for Early Breast Cancer Detection, Faria Rahman, Tasnime Mehejabin, Soniya Yeasmin, Manika Sarkar.

[8]. 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh 978-1-7281-7366-5/20/$31.00 © On Predicting and Analysing Breast Cancer using, Data Mining Approach.

[9]. Breast Cancer Detection Using Machine Learning Algorithms Sharma, Archit Aggarwal, Tanupriya Choudhury

[10]. Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques, AMIN UL HAQ, JIAN PING LI, ABDUS SABOOR, JALALUDDIN KHAN, SAMAD WALI, SULTAN AHMAD, (Member, IEEE), AMJAD ALI, GHUFRAN AHMAD KHAN, AND WANG ZHOU