www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 985-989

e-ISSN : 2583-1062

Impact Factor : 5.725

BREAST CANCER CLASSIFICATION USING MACHINE LEARNING

Mallamma C G[1], Sagar S T[2], Sagar S V[3], Nakul H M[4], Madhu P[5]

[1]Asst Professor, Dept. Of Cse, Sambhram Institute of Technology, Bengaluru-560097

[2,3,4,5]Eight Semester, Dept. Of Cse, Sambhram Institute of Technology, Bengaluru-560097, India

ABSTRACT

Breast cancer (BC) is the second most prevalent type of cancer among women leading to death, and its rate of mortality is very high. Its effects will be reduced if diagnosed early. BC' searly detection will greatly boost the prognosis and likelihood of recovery, as it may encourage prompt surgical care for patients. It is therefore vital to have a system enabling the healthcare industry to detect breast cancer quickly and accurately. Machine learning(ML) is widely used in breast cancer (BC) pattern classification duetoits advantages in modelling a critical feature detection from complex BC datasets. In this paper, we propose a system for automatic detection of BC diagnosis and prognosis using ensemble of classifiers. First, we review various machine learning (ML) algorithms and ensemble of different ML algorithms. We present an overview of ML algorithms including ANN, and ensemble of different classifiers for automatic BC diagnosis and prognosis detection. We also present and compare various ensemble models and other variants of tested ML based models with and without up-sampling technique on two benchmark datasets. We also studied the effects of using balanced class weight on prognosis dataset and compared its performance with others. The results showed that the ensemble method outperformed other state- of-the-art methods and achieved 98.83% accuracy. Because of high performance, the proposed system is of great importance to the medical industry and relevant research community. The comparison shows that the proposed method outperformed other state-of-the-art methods.

Key Words: Machine Learning, Who,

1. INTRODUCTION

Breast cancer is one of the most dangerous and prevalent cancers among women, causing the deaths of large numbers of women worldwide. Breast cancer accounts for 8.4% of diagnosed cancers and 6.6% of cancer-related deaths worldwide, according to a World Health Organization Breast cancer is more common in women with dense breasts, and there is a relationship between density and age, with younger women having denser breasts than older women. BC is the most frequently diagnosed cancer in women, accounting for approximately one in four newly diagnosed cancer cases. According to the World Cancer Research Fund International,Therefore detection of breast cancer at the earliest is vital to decrease the risk of developing cancer in other tissue cells and to carry out a proper treatment.There are two types of breast cancer(i.e., malignant and benign) invasive and non- invasive.The former is harmful, malignant, it has the ability to infect other organs/tissues, and is classified as cancerous. latter is non-invasive, not harmful, and does not spread to other organs/tissues. We presented an ensemble of machine learning-based methods for breast cancer diagnosis and prognosis using an ensemble of machine learning classifiers. We presented a comprehensive comparison of the performance of various machine learning and ensemble machine learning-based classifiers.We evaluated different sampling methods to address the class imbalance issue in our datasets. We demonstrated that the proposed method outperforms various state-of-the-art methods for the detection of breast cancer.

2. LITERATURE SURVEY

Many automatic systems for breast cancer classification have emerged in recent years; these systems use different approaches. Breast cancer categorization is a classification problem that requires the extraction of discriminatory fea tures and then classification. State-of-the-art strategies for breast cancer staging that have been proposed are discussedin the following paragraphs. Whitaker et al. [10] suggested a two-stage patch classi fication technique for mammography using two texture descriptors: ''Histogram of Oriented Texture (HOT)'' and ''Pass Band Discrete Cosine Transform (PB-DCT).'' In the first stage, mammogram patches are classified as nor mal or abnormal. The second stage uses a support vector machine (SVM) to classify aberrant mammographic regions as benign or malignant. Jothilakshmi and Raaza [13] devel oped a texture-based strategy to identify malignant and benign using multiple SVMs, with features retrieved using ''grey-level co-occurrence matrices (GLCM). In [14] pro posed a newapproach to classify benign and malignant breast masses. The approach converts two- dimensional contours of breastmasses on mammography into a one-dimensional signature. The one-dimensional signature is then segmented into subsections to extract local contour features. Finally, these features are fed to an SVM classifier. Laroussi et al. [39] Proposed two CAD systems for the classification of mammograms breast density for two and four ''BI-RADS'' classes consisting of features computed using different Law filters of varying lengths. The feature vectors are then fed to classifiers 'PNN,' 'NFC,' and 'SVM' to classify tissue densit.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 985-989

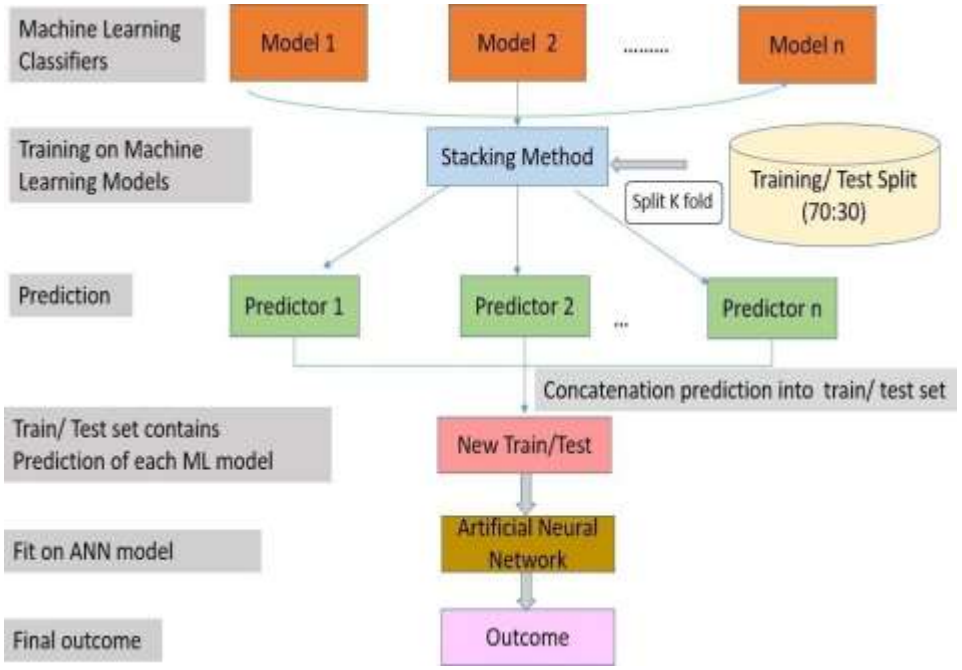
e-ISSN : 2583-1062

Impact Factor : 5.725

3. METHODOLOGY

This section presents the method used for an ensemble of ML classifiers. This architecture is composed of four different ML models. They are stacked and then further trained as an ensemble. After training, the ANN model is used for the outcome. The illustration of our proposed DL network is shown in Figure 4. The performance is compared with the several ML classifiers individually with and without up sampling techniques. We also compared the performance of the proposed ensemble model with other ensemble models.In this study, we design a classification framework by an ensemble of four ML-based classifiers named SVM, LR, NB, and DT. An ensemble model is stacked, and predictions are concatenated and then fed to the ANN model for final prediction. Each of the algorithmsused in our study is also briefly explained next section. The steps of the proposed model can be summarized below

1) We used machine learning based classifiers on a train- ing dataset
2) In the second step, the K-fold method retrieves the most common outcome from these classifiers.
3) In third step, we concatenated results from machine learning classifiers
4) New training dataset streamlined as a result
5) In this step, we input the new dataset into the default ANN
6) Result and evaluation of outputs



3.1 SUPPORT VECTOR MACHINE (SVM)

A supervised ML-based technique, SVM selects the mod- erate number of samples called support vectors and buildsa linear discriminant function. SVM solved the restriction of linear limits [40]. SVM can be considered a two-class data set that can be partitioned linearly to show a maximum hyperplane margin. The new samples are linearly fit orappear linearly separable in the high-level plane following the selec- tion of the appropriate mapping.

3.2 LOGISTIC REGRESSION (LR)

The LR method is created by replicating the posterior proba- bility of K groups across linear roles in x while ensuring they equal one and stay within the range [0, 1]. Logit shifts K -1, or log probabilities can be used to describe LR. Although the final group is used as the denominator in the odds ratio, the choice of the denominator is indeterminate because the counts are divided evenly. Since there is only one linear role when K 2, the style is direct. This technique is often used in biostatic tasks where binary responses are repeated.

3.3 NAIVE BAYES (NB)

Bayes theorem [42] is used to suggest the NB algorithm. The NB classifier can be revised in the following ways using Bayes' theorem and the exact procedures [43]. We conclude that there is a training set of instances T. There are group marks on these specimens. C1, C2 Ck are the names of the groups. Each specimen is an n-dimensional agent represented by the formula X x1, x2 xn. It states that X has n features since it has n dimensions. A specimen X is predicted to be a member of groupCi if the probability that group i depends on X is greater than the probability that each of the other groups depends on X, Using Bayes' Theorem P (Ci |X) is calculated as follows:P(Ci|X) = P(X |Ci)P(Ci)/P(X)

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 985-989

e-ISSN : 2583-1062

Impact Factor : 5.725

3.4 DECISION TREE (DT)

DT that begins with huge groupings of specimens within clearly defined categories [44]. Specimens are used for patterns that allow groups to be accurately characterized by com- bining nominal and numerical features. These markersare then represented as models, resulting in decision frameworks or sets of if-then processes that can be used to distinguish new samples, emphasizing making designs understandable and accurate. To determine the 'goodness' ofa test, the C4.5 cal- culus uses equations based on theoretical data; specifically, they choose the test that takes the most data from a collection of specimens while limiting themselves to evaluating a single characteristic.

3.5 ARTIFICIAL NEURAL NETWORK (ANN)

In the past decades, ANNs have been utilized by researchers, thus making them a relevant research area. Greatly, the net- work has enabled great success, especially in BC classifi- cation and early-stage prognosis [45]. ANN models usually have three layers: input, hidden, and output [31]. The layers comprise interconnected neurons with nonlinear switching activation functions to enhance nonlinear capacity. First, the input layer gets the data, then passes it to a hidden layer for analysis and returns the results to the output layer. Results shows are now displayed through the output layer. However, given the constraints, training an an ANN will likely require long informal chains of computing processes. There are three dense layers and two dropout levels in the ANN structure used in this study. The DNN, on the other hand, is made up of five dense layers and three dropout layers

4. EXPERIMENTAL RESULTS

In this section, we present the datasets used in this study and the experimental evaluations to demonstrate the usefulness of our proposed model. In this study, following previous studies, we used accuracy to evaluate the performance. Clas- sification results are analyzed using a 10-fold cross-validation technique.

4.1 DATASET DETAILS

The Breast cancer Wisconsin (Diagnosis)1 and Breast cancer Wisconsin (Prognosis)2 databases are used in this study

TABLE 1. Dataset distribution

Diagnosis		Prognosis	
malignant	benign	non-recur	recur
2	357	151	47

Wisconsin Breast Cancer (Diagnosis) contains 569 instances and 32 attributes (an ID and a target variable).Wisconsin Breast Cancer (Prognosis) contains 198 instances and 34 attributes (containing an ID and a target variable).The forecast dataset also had four missing attribute values,which were removed; furthermore, the forecast data set is considerably skewed, with 151 non-recurring and 47 recur- ring outcomes. Dataset distribution is given in Table 3. In the BC Wisconsin Diagnostic and Prognostic data sets, two additional strategies (algorithm approach and data approach) were implemented to solve the problem of an unbalanced classification problem. To start with, we used cost-sensitive learning or a misclassification penalty as a misclassification penalty while training the model to improve performance in minority classes

TABLE 2 : Comparison of ML classifiers. Average accuracy (%)

Machine Learning Models	Diagnosis	Prognosis
Decision Tree	91.22	76.26
Random Forest	97.07	76.26
Logistic Regression	98.00	75.24
Support Vector Machine	98.10	78.35

It can be observed from Table 4 that SVM outperforms all ML-based classifiers on both diagnosis BC with 98.10% accuracy and prognosis BC dataset with 78.35% accuracy. In contrast, the worst classifier on the diagnosis dataset is DTwith 91.22% accuracy and NB on the prognosis dataset with70.71%. For DL-based classifiers, ANN performs well on both datasets, with 98.24% accuracy on the diagnosis datasetand 90.22% accuracy on the prognosis dataset compared to DNN.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 985-989

e-ISSN : 2583-1062

Impact Factor : 5.725

TABLE 3: Ensemble of ML classifiers with ANN with and without samplying

Model	Diagnosis		Prognosis	
	Without Sampling	With Sampling	Without Sampling	With Sampling
(SVM+LR+NB+DT)+ANN	97.67%	98.83%	81.35%	84.70%
(SVM+LR+NB+RF)+ANN	97.07%	98.24%	83.05%	88.13%
(SVM+LR+RF+DT)+ANN	97.66%	98.24%	83.00%	84.74%
(SVM+LR+RF+NB) + ANN	97.07%	98.24%	83.15%	88.33%
(SVM+LR+RF) + ANN	95.91%	98.14%	81.36%	77.96%
(SVM+LR) + ANN	96.46%	96.46%	76.27%	76.27%

Figure 1: Illustration of confusion matrix (Diagnosis).

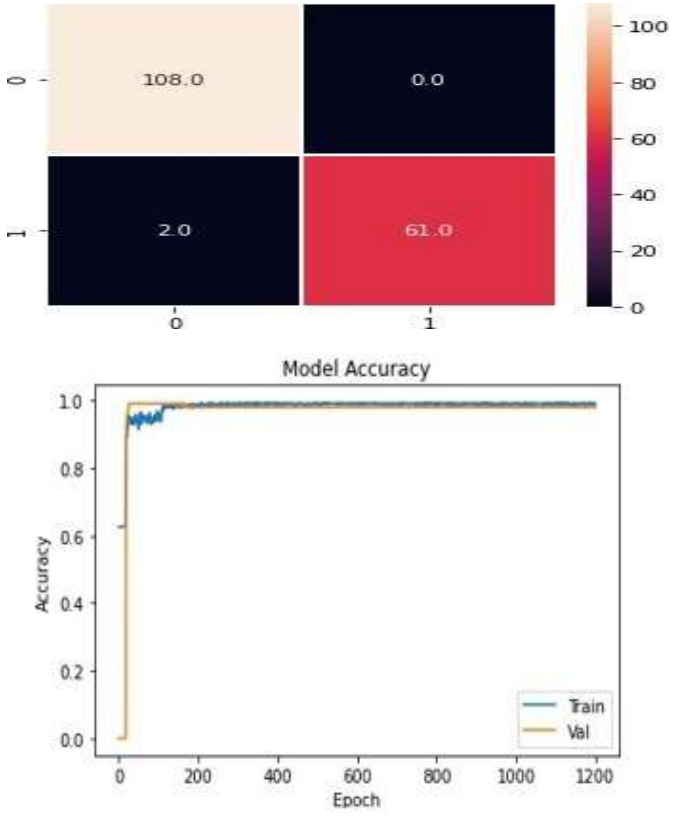


Figure 2: Illustration of train/test accuracy (Diagnosis)

Comparing all the different ensemble models in Table 3 shows which model performed best for each dataset. The best ensemble model is the ensemble of (SVM LR NB DT) in both cases (without 97.67% and with upsam- pling 98.83%). In contrast, the worst-performing combina- tion is (SVM LR RF) in both cases (95.91% for without sampling and 98.14%for upsampling) on the diagnosis dataset. For prognosis, the best ensemble model combines (SVM LR RF NB) in both cases (without 83.15% and with upsampling 88.33%). In contrast, the worst-performing combination is (SVM LR) in both cases (76.27% for without and 76.27% with up-sampling) on prognosis. The increment of 1.16% was observed on diagnosis and 5.18% on the prog- nosis dataset when the upsampling technique was used. The confusion matrix and train/test accuracy of best-performing ensemble classifiers can be seen in Figure 1 and Figure 2, respectively.

5. ANALYSIS

We also analyzed the effects of applying balanced class weights with sampling and measured the performance whenWe observed that performance increased substantially for all tested combinations of clas sifiers when compared with upsampling on the prognosis dataset. We also note that the confusion matrices below show that when K is 5 instead of 10, the model (SVM + LR + RF + DT) trained on the forecast outperforms.

6. CONCLUSION

We proposed a method for breast cancer diagnosis and prog- nosis using machine learning techniques in this research. Benchmark datasets are used for the experiments. Classifiers based on machine learning and deep learninghave shown their exceptional potential to increase classification and pre- diction 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, (SVM LR NB DT) performs well without and with upsampling on the diagnosis dataset, whereas (SVM LR RF NB) outper- forms 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 upsampling technique as compared to without, and the upsampling technique is used individually. The perfor- mance was also analyzed 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

7. REFERENCES

[1] WHO | World Health Organization.” [Online]. Available: https://www.who.int/. [Accessed: 28-Jun-2021]

[2] D. Bardou, K. Zhang, and S. M. Ahmad, “Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks,” IEEE Access, vol. 6, pp. 24680–24693, 2018.

[3] Priyanka and K. Sanjeev, “A review paper on breast cancer detection using deep learning,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021.

[4] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran, and K. Ur Rehman, “A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities,” IEEE Access, vol. 8, pp. 165779–165809, 2020.

[5] V. Lahoura, H. Singh, A. Aggarwal, B. Sharma, M. A. Mohammed, R. Damaševičius, S. Kadry, and K. Cengiz, Cloud computing-based framework for breast cancer diagnosis using extreme learning machine,” Diagnostics, vol.11, no. 2, pp. 1–19, 2021.

[6] S. Beura, B. Majhi, and R. Dash, ‘‘Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer,’’ Neurocomputing, vol. 154, pp. 1–14, Apr. 2015.

[7] R. AlamKhan, N. Ahmad, and N. Minallah, ‘‘Classification and regres- sion analysis of the prognostic breast cancer using generation optimizing algorithms,’’ Int. J. Comput. Appl., vol. 68, no. 25, pp. 42–47, Apr. 2013.

[8] R. Nisbet, J. Elder, and G. D. Miner, Handbook of Statistical Analysis and Data Mining Applications. New York, NY, USA: Academic, 2009.

[9] Onan, ‘‘A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for auto- mated diagnosis of breast cancer,’’ Expert Syst. Appl., vol. 42, no. 20, pp. 6844–6852, 2015.

[10] Y.-Q. Liu, C. Wang, and L. Zhang, ‘‘Decision tree based predictive models for breast cancer survivability on imbalanced data,’’ in Proc. 3rd Int. Conf. Bioinf. Biomed. Eng., Jun. 2009, pp. 1–4.

[11] J. R. Quinlan, ‘‘Improved use of continuous attributes in C4.5,’’ J. Artif. Intell. Res., vol. 4, pp. 77–90, Mar. 1996.

[12] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, ‘‘A support vector machine- based ensemble algorithm for breastcancer diagnosis,’’ Eur. J. Oper. Res., vol. 267, no. 2, pp. 687–699, Jun. 2018.