Multimedia Tools and Applications (2023) 82:18059–18076
https://doi.org/10.1007/s11042-022-14265-1

Breast cancer diagnosis using Stochastic Self-Organizing Map and Enlarge C4.5

Arvind Jaiswal[1] · Rajeev Kumar[2]

Received: 20 July 2021 / Revised: 7 November 2022 / Accepted: 8 November 2022 /
Published online: 1 December 2022

Abstract

Timely and accurate Breast Cancer (BC) prediction allows healthcare providers and doctors to take suitable decisions to treat the patients. Thus, this study employed a strategy based on Deep Learning (DL) to diagnose BC. The study intends to cluster the BC data by the proposed Stochastic Self-Organizing Map (SOM) as it has the ability to process complex data. On the other hand, Enlarge C4.5 (E-C4.5) algorithm is introduced to predict the BC cases based on the clustered outcomes. The BC dataset is loaded and pre-processing is performed where dimensionality reduction is executed to select only the relevant features for clustering. This process eases the clustering. Then, clustering is undertaken by the proposed Stochastic Self-Organizing Map. Here, all the identical data are grouped as clusters which make it easy for prediction. Followed by this, Breast Cancer is predicted by the proposed Enlarge-C4.5 algorithm. After this, the predicted results are analysed by comparative analysis through the four standard performance metrics. This analysis is significant as it shows the degree to which the introduced techniques are effective than the existing techniques. The histogram, correlation map, confusion matrix and clustering results are also discussed clearly. The analytical outcomes explore that the proposed methods are effective than the conventional methods as the proposed method shows a high accuracy rate, precision rate, recall rate and F1-score rate. The misinterpretation rate is also found to be minimum on implementing the proposed method, which is confirmed through the confusion matrix. The proposed method also determines the malignant and benign counts.

Keywords Breast cancer · Prediction · Dimensionality reduction · Stochastic Self-organizing Map and Enlarge C4.5 algorithm

✉ Rajeev Kumar
rajeev2009mca@gmail.com

Arvind Jaiswal
arvindjsir@gmail.com

[1] Acropolis Institute of Technology and Research, Indore, M.P., India

[2] G.L. Bajaj Institute of Technology & Management, Greater Noida, U.P., India

1 Introduction

Breast Cancer (BC) has become the main cause of death among women. Accessing the healthcare datasets and analysis promotes the investigators to employ the research in unknown pattern extraction from the healthcare datasets. The study [25] aimed to propose an analytical data model that can support a clear understanding of BC survivability in missing data presence, affording clear visions into factors related to the survivability of patients and forming patient's cohorts that share identical properties. Separating patients into clusters enhanced the accuracy of survival prediction that rely on MLP and explored intricate conditions. This affects the prediction accuracy. Thus, the proposed strategy that relies on unsupervised learning techniques enhanced the understanding, thereby helping to find patterns related to patient survivability. The analytical results can be utilized to perform patient data segmentation into subsets or clusters that share common survivability and variable values. The accuracy in predicting survival rate is enhanced when MLP used recognized patient cohorts instead of raw historical data. Variable value analysis in individual cohort affords better visions into survivability of specific subgroup of BC patients [24]. The proposed strategy can be employed to classify other databases like Mammographic Image Analysis Society (MIAS). This has to be implemented in the near future. Experiments have been conducted by utilizing different classifiers on Wisconsin Breast Cancer Dataset (WBCD) [18]. The proposed method is analysed by taking into account the predicted and actual classification. It has been found that the KNN classifier produces maximum classification accuracy when utilized with more predictive variables. Future work concentrates on exploring many dataset values, thereby affording fascinating results. This study aided in creating a highly reliable and effective diagnostic system for disease prediction, which will subsidize towards evolving effective healthcare systems, thereby minimizing time, mortality rate and overall cost. Additionally, the article [17] proposed twelve Machine Learning (ML) methods for diagnosing BC. Promising outcomes have been attained with an accuracy rate of more than 94%. Yet, the study has to be extended by employing Deep Learning (DL) and Artificial Neural Network (ANN) for developing the predictive model with a large unstructured dataset. This extension is needed to enhance the prediction accuracy [7]. Thus, to enhance the prediction accuracy, the present study proposed Stochastic Self-Organizing Maps (SOM) and Enlarge C4.5 algorithm to predict BC based on DL. Instead of segmentation and classification, the study employed a clustering technique to increase the accuracy rate in BC prediction.

The major contributions of this study are listed below.

- To perform clustering by the proposed Self-Organizing Maps (SOM) for grouping identical data together and enhance the prediction.
- To predict the malignant and benign Breast Cancer (BC) data based on the clustering using the proposed Enlarge C4.5 (E-C4.5) algorithm.
- To analyse the prediction results of the proposed system by comparing it with the existing methods in terms of performance metrics such as accuracy, F1-score, recall and precision.

1.1 Paper organization

The paper is organized in the following way. Section 1 discusses the fundamental ideas related to BC prediction. Followed by this, the various strategies employed by the traditional systems for diagnosing BC is explained in Section 2. After this, the overall proposed methodologies relevant to this context are described in Section 3. The results obtained after the implementation of the proposed methods are discussed in Section 4. At last, the overall summarization of this study is presented in Section 5.

2 Review of existing work

Breast-cancer is a cancer formed in the cells of breasts. BC occurs in both women and men, but most commonly in women. Breast cancer is also called as a metastatic cancer and it can transfer to a distant organ namely the bone, lung, liver and brain, are incurability. There are numerous risk factor like sex, family history, age, estragon, unhealthy life style and gene mutations, can increases the possibility of developing breast-cancer. Breast Cancer (BC) is a typical disease related to poor prediction. Thus, there exists an urgent requirement to develop fast and efficient computational techniques for the prediction of BC. The study [21] examined the Deep Neural Network (DNN) model's performance on the classification task associated with the detection of BC. The proposed strategy intended for BC tumour classification as malignant or benign. Various tumour subclasses have also been predicted, like Lobular Carcinoma, Fibroadenoma and so on. The empirical outcomes on the histopathological images by the use of the Break his dataset revealed that the proposed model accomplished effective performance with an accuracy of 95.4% in the classification of multiclass BC in comparison to traditional techniques. The performance of the proposed method can be enhanced when more data can be afforded by the use of large datasets. Similarly, this paper [6] examined the Break his dataset from four varied viewpoints, namely Magnification Specific Multi-category (MSM), Magnification Specific Binary (MSB), Magnification Independent Multi-category (MIM) and Magnification Independent Binary (MIB). The top reformulation has been identified from a practical and clinical standpoint by examining varied aspects utilized to describe the proposed taxonomy. It has been explored that the MIM reformulation is suitable to solve this issue. On the other hand, the current organization and state of the dataset (Break his) constructing an automatic system by the use of Convolutional Neural Network (CNN)failed to provide efficient accuracy. Moreover, the study has to examine the ability of additional Deep Learning (DL) models like Deep Belief Networks (DBNs). In accordance with this, the article [3] examined a DL based CNN strategy for short-term risk prediction to develop BC by the use of general screening mammograms. The obtained results exhibited that the proposed GoogLeNet – LDA model performed better than the GoogLeNet model. It has also been found that these two models performed well when compared to Mammographic Breast Density (MBD). This introductory study represented the promise and feasibility of employing DL to improvise BC risk assessment, guaranteeing larger multi-centre research to further assess the proposed model. Though it is advantageous, the study is in its initial stage and has to be improved. Likewise, the potential of the DL model has been evaluated to differentiate malignant and benign breast lesions by the use of MRI [12]. It characterized the various BC subtypes. Efficient outcomes have been obtained. Yet, the model's accuracy has to be enhanced by using multi-parametric MRI and large databases. On the contrary, the study [20] trained and then assessed a more deeper network termed Deep Dilated Residual Network (DD-ResNet) for segmentation in an automatic way for planning Computed Tomography (CT) of Breast Cancer (BC). The quantitative outcomes exhibited a better accuracy rate with satisfactory time consumption. Predicting the time of survival for human BC is significant.

Thus, a multi-modal DNN has been integrated with multi-dimensional data to find this survival time [27]. The proposed method can also be employed for other diseases, which is yet to be done [15].

In addition, Transfer Learning (TL) has been applied to solve the shortcomings in traditional systems to detect and then classify the BC tumours [16]. The features are extracted from the BC images by the use of VGGNet, ResNet and GoogLeNet. This is combined with TL for enhancing the classification accuracy. Data augmentation has also been introduced to improvise the dataset size so as to enhance the CNN structure's efficacy. The results showed that the suggested framework affords excellent outcomes in terms of accuracy. CNN features and handcrafted features have to be utilized to enhance the accuracy in classification. Similarly, a DL method has been employed that fine-tuned Bi-directional Encoder Representations from Transformers) for BC attribute and concept extraction [32]. The outcomes confirmed the superior performance of the proposed method in comparison to the conventional Machine Learning (ML) algorithms. This assists its utilization in wide Extraction Tasks and Named Entity Recognition (NER) in the medical field. The study also assists in affording high performance in other medical areas [31]. Additionally, the paper [23] proposed the utilization of an effective DL framework for nucleus and cell membrane detection, segmentation followed by classification as well as scoring by the use of HER2-stained ImmunoHistoChemical (IHC) images. The outcomes represented that the introduced method is feasible to quantify and score the status of HER2 using BC-IHC images. On the contrary, Graph Convolutional Network (GCN) and CNN have been combined to produce highly accurate diagnosis by utilizing breast mammograms [33]. Initially, the study designed Net-0, which is a base network. Subsequently, it incorporated two enhancement methodologies to attain an enhanced Artificial Intelligence (AI) model named Net-1. Followed by this, Net-2 has been developed by using Rank based Stochastic Pooling (RSP) for replacing Net-1's conventional pooling. Then, Net-0, Net-2 and Net-1 have been integrated with the proposed GCN to attain Net-3, Net-5 and Net-4. The proposed methods are analyzed to evaluate their efficiency and effectiveness for BC prediction. The results revealed that Net-5 affords effective outcomes than the other six networks proposed in this study. On the other hand, Net-5 was also effective than the traditional methods. However, the proposed method has to be applied to huge datasets and test the efficacy of the proposed model. The model has to be tested on various sources of breast mammogram images with different resolutions. The CNN and GCN has to be assessed with other combination method to improvise its efficiency.

Different techniques have been applied to predict BC. The article [11] built eight Artificial Neural Network (ANN) models by utilizing three clustering techniques such as K-means, Fuzzy K-means and Random pick. In addition, two techniques have been employed to calculate the weights between Normal equations, Gradient Descent and Multi-Layer Perceptron alternatives (three layers of basic MLP and four layers of Deep MLP). Subsequently, choosing optimal Radial Basis Function Network (RBFN) and MLP alternatives with respect to F-measure and accuracy. This work intends to utilize and evaluate other local and global interpretation methods since this will be fascinating to improve the model's trust and gain the attention of oncologists to make use of these black-box models. Besides, enhancing the tuning of MLP parameter might occur to accomplish effective performance outcomes. Additionally, the paper [12, 29] assessed four classification models that comprised of Support Vector Machines (MLP), Naïve Bayes (NB), Decision Tree (DT) and K-Nearest Neighbour

(K-NN) by Feature Selection (FS) at varied threshold levels for model training so as to classify the two kinds of BC. The analytical outcomes explored that the proposed algorithms were able to classify BC accurately as Triple Negative BC (TNBC) or non-TNBC. Though all the algorithms exhibited effective results, SVM was outstanding than the other three algorithms. Yet, further research has to be carried out to examine the ML algorithm efficacy in BC classification [8, 13]. On the other hand, Decision Tree-based Ensemble Learning (DT-EL) has been employed for BC classification. In accordance with the statistical analysis, the introduced techniques were able to perform BC classification accurately [30]. In addition, the metabolic network has been assessed that rely on Divergence Encoding (DE) [4]. The study also used this approach for BC. This strategy can also be extended to other types of cancers. The proposed method has the capacity to afford mechanistic visions into metabolic dependencies that are cancer-specific and permit the detection of probable medicine targets for individual patients. Similarly, the article [1, 5] concentrated to examine Data Mining (DM) methods to bring effective BC prediction. The BC prediction initiates with the examination of traditional BC with respect to diagnosis. The hypothesis has been determined from the literature survey, which was the basis for the study. The risk analysis of BC, survivability and recurrence prediction has to be implemented in future. Furthermore, a clustering algorithm named Intuitionistic Fuzzy Soft Set – IFSS has been proposed for segmenting the mammogram image [9]. The proposed model aided to find the lumps, nodules, lesions or other kind of breast abnormalities in the initial stage of BC. Moreover, the bi-clustering algorithm has to be employed under the Fuzzy Soft Framework (FSF) for segmenting the mammogram so as to identify BC in its initial stage itself [19, 22].

3 Proposed methodology

Predicting BC has become significant. The study employed various processes to diagnose BC. Clustering is performed by the proposed Stochastic Self-Organizing Maps (SOM). The major advantage of proposing SOM is it has the ability to interpret data easily. The dimensionality reduction helps it to find similar data, thereby performs effective clustering. In addition, Enlarge C4.5 (E-C4.5) is proposed to predict the BC data as malignant or benign that rely on clustering. The E-C4.5 algorithm has the ability to handle continuous and discrete data, thereby predicts the data effectively. Various processes are involved to accomplish the BC data prediction. It is given in Fig. 1.

Initially, the Breast Cancer dataset is taken as input and dimensionality is reduced in pre-processing stage to select the relevant features. After this, clustering is performed by the proposed Stochastic SOM. In this process, all the identical data are grouped as clusters so as to make the prediction easier. The E-C4.5 is applied to predict the BC data. Finally, the prediction results are analyzed to evaluate the efficiency of the proposed techniques. Performance metrics taken into consideration for analysis include recall, accuracy, F1-score and precision.

3.1 Stochastic Self-Organizing Maps

Self-Organizing Map (SOM) is the well-renowned Neural Network (NN) models. It pertains to the competitive learning network category. This algorithm relies on unsupervised learning.

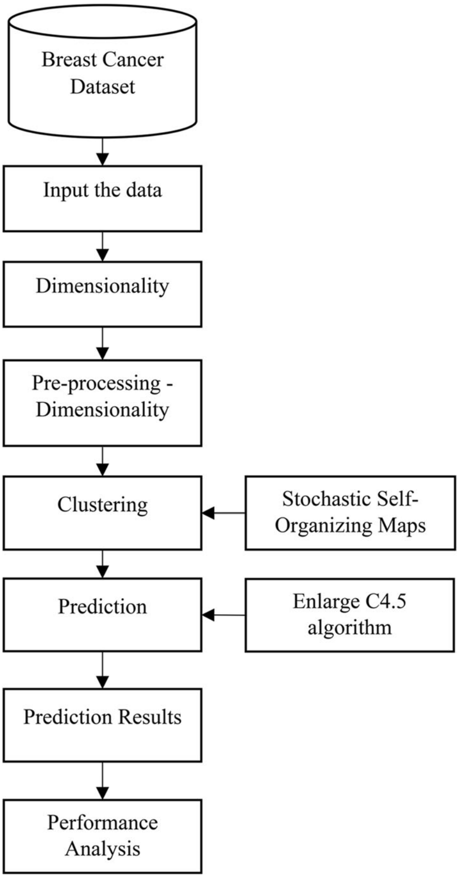

Fig. 1 Overall view of the proposed system

Hence, human interference is not required during learning. Additionally, only little information has to be known regarding the input data characteristics. The study uses SOM to cluster data without the knowledge of input data's class memberships. Features that correspond to the issue are detected by using SOM and thus, it is also called as Self-Organizing Feature Map (SOFM). This algorithm has numerous applications and is applied in various areas for complex data processing like time series, nominal data, kernel data or categorical data and has the capacity to generalize, which means the network can characterize or recognize new inputs (the inputs that it has not come across before). The methodology also affords a topology conserving a mapping from high-dimensional space to the neurons. The neurons or mapping units typically build a Two-Dimensional (2D) lattice. This mapping occurs from high-dimensional space to a plane. The topology conserving property indicates that the mapping conserves the associative distance amongst the points. Furthermore, points in the input space that are at a minimum distance with one another are mapped to the neighbouring neurons in SOM. Thus, this particular algorithm can serve as an efficient analysing tool for clustering high-dimensional data. The pseudo code of this algorithm is discussed below.

Algorithm I Stochastic Self-Organizing Maps

Step 1: Initialize the reference vectors of Stochastic SOM to random values.

Step 2: Fix the initial radius corresponding to the neighbourhood area.

Step 3: Compute the distance D_i between the present input vector X_i of N-dimension and each of the neuron's reference vectors in the mapping array in accordance with the Euclidean distance (ED) function given by

$$D_i \sum_{ii=1}^{N_i} \left(X_{ii} - W_{jii}\right)^2$$

Here X_{ii} is theX_i input vector'siithelement and $W_{jii}$is the iith element of the W_{ji} reference vector for the ji neuron.

Step 4: Find the top similar neuron φ that possess minimal distance.

Step 5: Update the neuron's reference vectors in the neighbourhood area around the top similar neuron φ. The reference vector is updated and is given by

$$\triangleq W_{jii} = n\left(X_{ii} - W_{jii}\right)$$

Here n indicates the Learning Rate Parameter (LRP).

Step 6: Reiterate the steps two to four for the subsequent input vector till the given criteria are accomplished. During step 4, the LRP and neighbourhood area n is monotonically reduced over time.

Step 7: To cluster the BC data using the proposed Stochastic SOM, the conventional labelling technique is modified. Each neuron in the stochastic SOM will yield a ratio value. This output neuron value is given by $(0 \leq r_i \leq 1)$

$$\text{Here } r_i = \begin{cases} 0.5 & \\ n_{mi} \backslash n_{i_b} + n_{i_m} & \text{else} \end{cases}$$

n_{i_b} Denotes the count of benign cases and n_{mi} represents the count of malignant cases in the BC dataset. This is clustered by the proposed stochastic SOM.

At first, the reference vectors of Stochastic SOM are initialized with random values and the initial radius of the neighbourhood area is also fixed. Then, the distance between the present input vector and the individual neuron's reference vectors are calculated as per step 3 to find the similar neurons with minimum distance and the reference vectors of neurons are updated according to step 5. Steps 2 to 4 are iterated till the stop condition is met. Finally, the BC data is clustered through the proposed SOM by modifying the traditional labelling method by step 7 to obtain the groups of identical BC data. This process helps in increasing the BC prediction rate.

3.2 Enlarge C4.5 algorithm (E-C4.5)

The Enlarge C4.5 is employed for BC prediction. This algorithm is utilized as a Decision Tree (DT) to make decisions that rely on specific data samples that comprise of multi-variants or univariante predictors. The proposed E-C4.5 is the practical and widely utilized technique for inductive interpretation. It is a technique to estimate functions of discrete values that have the ability to learn disjunctive expressions and is strong to noisy data. Here, the learned function is denoted by a DT. It can also be denoted by if-then rules for enhancing human readability. In addition, these learning techniques are successfully employed to diagnose various medical cases. It is a one-step look ahead, heuristic and non-backtracking search algorithm.

The pseudo code of this algorithm is given below.

Algorithm II Enlarge C4.5

```
$\text{Decision}_{\text{Tree}}$ (eg)
Prune $(\text{Tree}_{\text{generation}}(\text{eg}))$
$\text{Tree}_{\text{generation}}(\text{eg}) =$
IF $\text{end}_{\text{condition}}(\text{eg})$
Then leaf $(\text{major}_{\text{class}}(\text{eg}))$
Else
Let
$\text{Top}_{\text{test}} = \text{selection}_{\text{function}}$ (eg)
In
For each $\text{Top}_{\text{test}}$ value
Let $\text{subtree}_{\text{v}_\text{i}} = \text{Tree}_{\text{Generation}}(\text{e}_\text{i}\text{eg} \backslash \text{e}_\text{i}\text{Top}_{\text{test}} = \text{v})$
In Node $(\text{Top}_{\text{test}}, \text{Subtree}_{\text{v}_\text{i}})$
```

The functioning of the proposed E-C4.5 algorithm is given as below.

- Choose the attribute, formulate the attribute's logical test.
- Branch on the individual test outcome, move the training data (a subset of instances), satisfying the child node outcome.
- Run iteratively on the individual child node.
- The end rule indicates the exact time for leaf node declaration.

Definitions that utilized training of the proposed E-C4.5 are explained here. As per the above Algorithm II, the pruning algorithm avoids overfitting and the selection function is utilized to partition the training data. Finally, the end condition finds when to terminate the partitioning. Hence, this algorithm predicts the malignant and benign BC cases.

4 Results and discussion

The proposed method is implemented for BC data and the obtained results are discussed in this section. Furthermore, the experimental outcomes are explained. The prediction results of the proposed methodology are analysed by using performance metrics such as accuracy, precision, F1-score and recall. The analytical outcomes are also discussed here.

4.1 Dataset description

Wisconsin Diagnostic Breast Cancer dataset is utilized in this study. This dataset consists of thirty-two features of five hundred and sixty-nine subjects. The thirty-two features comprises thirty real tumour features, ID number as well as the class label. This represents that the individual subject possesses a malignant or benign tumour. In this study, the dataset is improved. The dataset is obtained from Breast Cancer Wisconsin (Diagnostic) Data Set, General Surgery Dept. University of Wisconsin, Clinical Sciences Centre, Madison [26, 28].

4.2 Performance metrics

The performance of the proposed system is analysed in terms of precision, accuracy, f1-score and recall. It is clearly explained below.

A. **Accuracy**

This metric assesses the ability of the system to predict the BC data as benign or malignant. It is given as per Eq. 1.

$$\mathrm{Accuracy} = (\mathrm{TN} + \mathrm{TP})/(\mathrm{TN} + \mathrm{FN} + \mathrm{TP} + \mathrm{FP}) \tag{1}$$

Here TN is True Negative, TP is True Positive, FN is False Negative and FP is False Positive.

B. **Precision**

It explores the extent to which a particular process replicates identical values and is given by Eq. 2.

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{2}$$

Here TP is True Positive and FP is False Positive.

C. **F1-score**

F1-score also known as F-measure is the harmonic mean of Recall and Precision. It affords effective measurement and is given by Eq. 3.

$$\mathrm{F1} - \mathrm{Score} = 2 * (\mathrm{R} * \mathrm{P})/(\mathrm{R} + \mathrm{P}) \tag{3}$$

Here R is Recall and P is Precision.

D. **Recall**

This metric quantifies the count of accurate positive predictions out of each positive predictions that might be made. It is given by Eq. 4.

$$\mathrm{Recall} = \mathrm{TP}/(\text{Predicted outcomes}) \tag{4}$$

4.3 Experimental outcomes

The study used Stochastic SOM and E-C4.5 to cluster the identical BC data into groups and then predict it accordingly. In this section, the histogram, count of malignant and benign BC data, confusion matrix, correlation map and clustering results are discussed. The outcomes obtained after clustering is shown in Fig. 2.

The proposed Stochastic SOM clusters the BC data as malignant or benign. Here the red circle in the boundary box denotes the malignant BC cases, while the green squared boundary

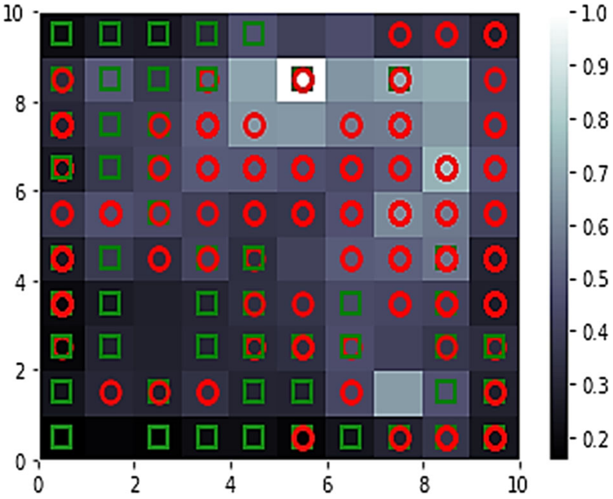

Fig. 2 Clustering results

box denotes the benign BC cases. Additionally, the histogram for the two kinds of BC (malignant, benign) is given in Fig. 3.

In the above Fig. 3, the radius mean values and frequencies are considered. As per the obtained histogram, the benign rate is more than the malignant rate. As it is randomly shown, the actual benign and malignant cases are also computed, which is shown in Fig. 4.

The implementation of the proposed method helps to find the benign (B) and malignant (M) count. It is found that the malignant cases are above 400. Whereas benign cases are found to be above 700. This gives a clear view of the number of BC cases to find cancerous and non-cancerous patients. In addition, a confusion matrix is also computed, as shown in Fig. 5, to find the correct and misinterpretations.

The confusion matrix reveals that 152 cases are correctly predicted as malignant and there is no misinterpretation in this case. On the other hand, 2 are misinterpreted as benign and 74

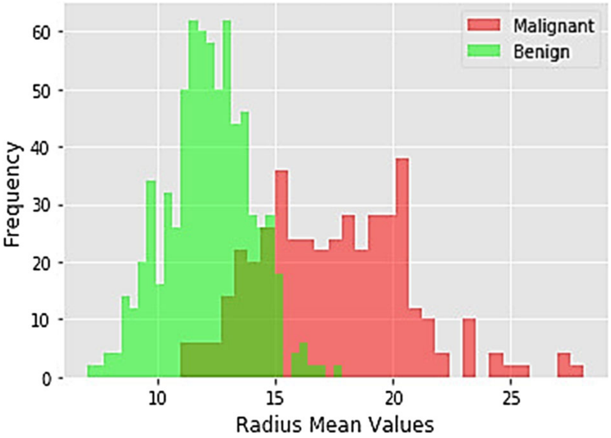

Fig. 3 Radius mean histogram for malignant and benign tumours

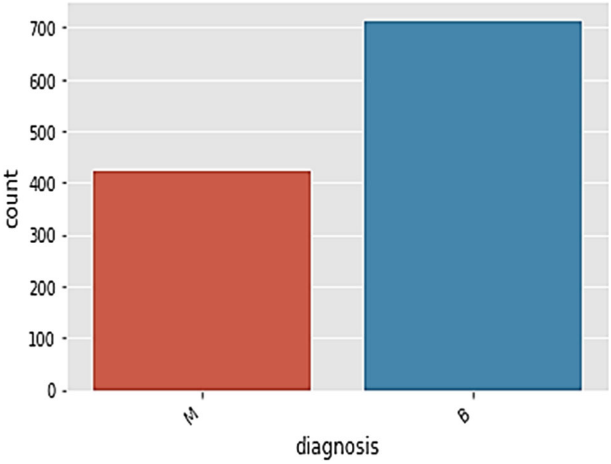

Fig. 4 Count of benign and malignant BC cases

are correctly identified as benign. As the misinterpretation rate is minimum than the correct prediction, it is clear that the proposed methodologies are effective.

Furthermore, a correlation map is also attained, as shown in Fig. 6. Here, various features are taken in both the X and Y axis, which also assists to explore effective, relevant features. The light colours in the correlation map represent the positive correlation. Here the dark colours represent a negative correlation.

It is found that the diagonal has light colours in the correlation map (Fig. 6), which indicates that feature selection is performed effectively. Hence, it is concluded that the proposed methodologies explore effective results on execution to accurately predict the BC only with slight misinterpretation.

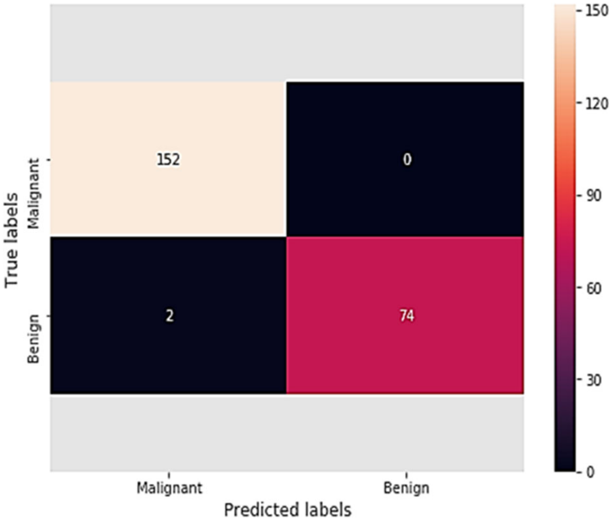

Fig. 5 Confusion matrix

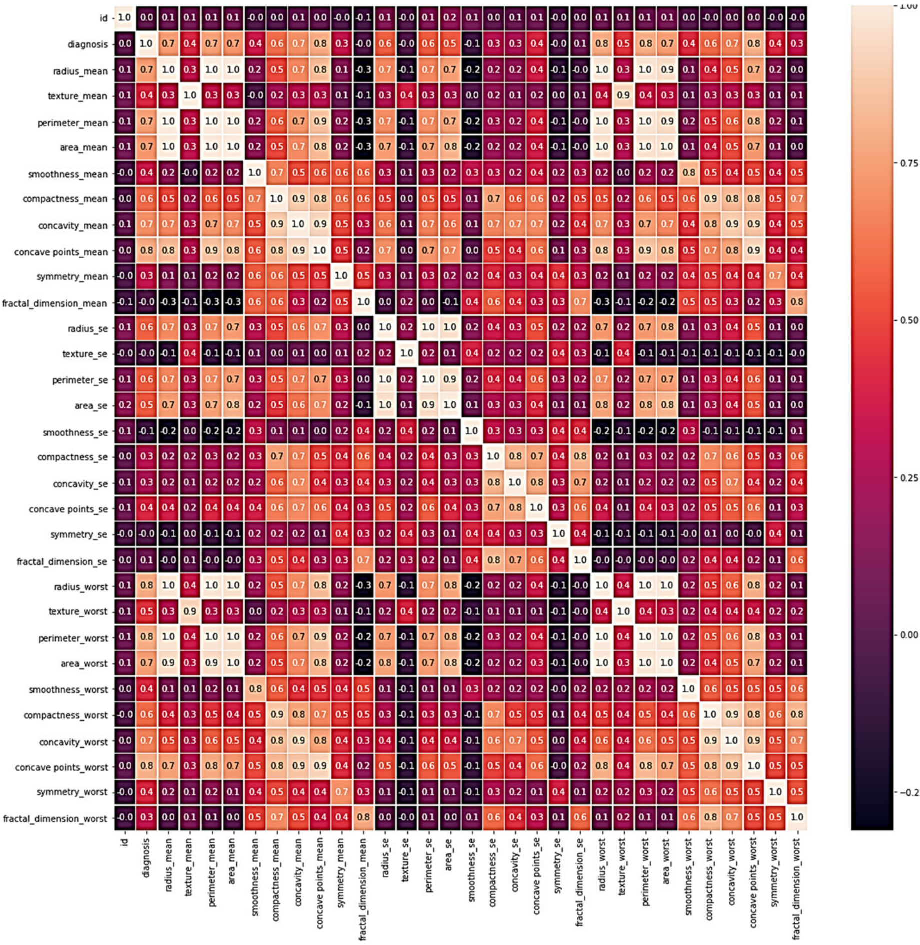

Fig. 6 Correlation map

4.4 Performance analysis of the predicted results

The performance of the proposed methodology is assessed by comparing with traditional methods with respect to the accuracy, recall, precision and F1-score. This analysis aids to find the degree to which the introduced techniques are superior in performance than the existing methods. Random Forest (RF), eXtreme Gradient Boosting (XGBoost), IGSAGAW, Cost Sensitive Support Vector Machine, GAW + cost sensitive SVM, IGSAGAW + cost sensitive SVM, Stacked Sparse AutoEncoders-SoftMax Regression (SSAE-SM), Feature Ensemble SSAE-SM, K-Nearest Neighbour, Logistic Regression (L_G), DT, SVM and DL_ANN are the various existing methods taken for comparative analysis. At first, RF and XGBoost are compared with the proposed method with respect to the four performance metrics, as shown in Table 1.

The results show that the accuracy rate of RF is 97.07% and the accuracy rate of XGBoost is 98.53%. Here the proposed method is showing a high accuracy rate of 99.12%. Similarly,

Table 1 Analysis of the proposed and existing methods [2]

Metrics	RF	XGBoost	Proposed
Accuracy	97.07%	98.53%	99.12%
Precision	97%	98%	99%
Recall	97%	99%	99%
F1-score	97%	98%	99%

the precision, F1-score and recall rate is also found to be higher for the proposed than the existing methods. It is graphically shown in Fig. 7.

From Fig. 7, it is clear that the proposed method is effective in terms of all the four mentioned metrics, thereby enhancing the system efficiency. Additionally, the proposed method is compared with other traditional algorithms to find the efficiency in prediction accuracy. It is given in Table 2.

Though all the methods afford an effective accuracy rate, the proposed technique exhibits a high accuracy rate of 99.122%. As the proposed E-C4.5 has the ability to learn disjunctive expressions and SOM performs complex data processing and effective clustering, the implementation of these algorithms results in a high accuracy rate. It is graphically shown in Fig. 8.

From the Fig. 8, it is clear that the proposed method is efficient in terms of accuracy than the other traditional methods considered here. Further, the proposed method is compared with many other conventional techniques to evaluate the extent to which the introduced methodologies are effective and efficient. It is shown in Table 3.

Here KNN, L_G, DT, SVM, RF and DL_ANN are considered for comparison. According to the comparative analysis, the existing KNN, L_G and DT show accuracy at a rate of 95.8%. Similarly, RF and SVM show similar accuracy rate of 97.2%. ere the existing DL_ANN

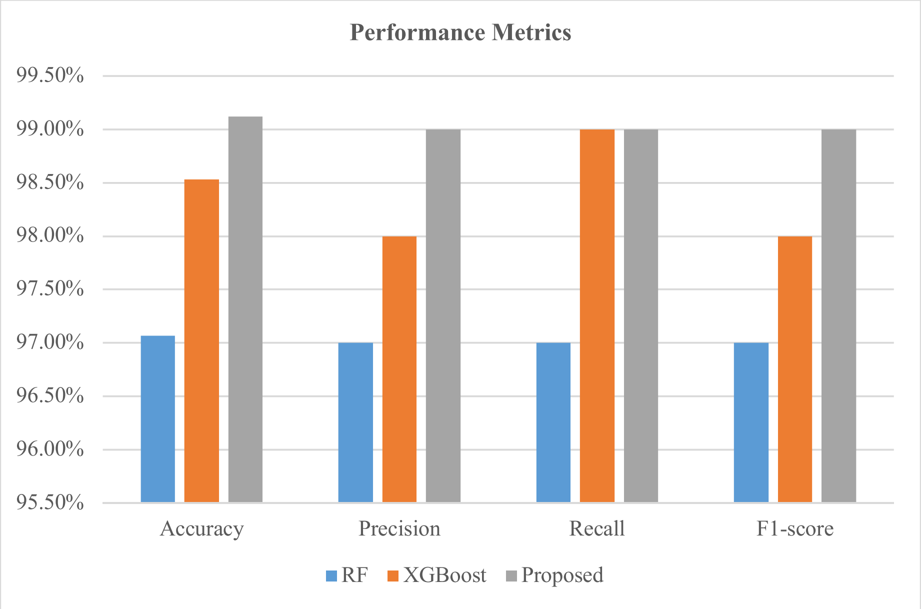

Fig. 7 Performance analysis of the existing [2] and proposed method with respect to various performance matrices

Table 2 Analysis of the proposed and existing methods [14] in terms of accuracy

Method	Accuracy(%)
IGSAGAW	95.4
Cost sensitive SVM	92.6
GAW+cost sensitive SVM	94.5
IGSAGAW+cost sensitive SVM	95.7
SSAE-SM	98.25
FE-SSAE-SM	98.6
Proposed	99.122

explores 98.24% as accuracy rate. Moreover, the proposed method shows a high accuracy rate of 99.12% in this analysis when compared to these mentioned existing algorithms. Likewise, the recall, precision and F1-score of the existing methods are found to be low than the proposed method. It is graphically shown in Fig. 9.

After the comparative analysis, it is found that the proposed method is effective than KNN, L_G, DR, SVM, RF and DL_ANN with accuracy at a rate of 99.12%, 98.5% of precision, recall and F1score. As the performance of the proposed method is outstanding than the other existing methods [10], the proposed Stochastic SOM and E-C4.5 algorithm are highly suitable for BC prognosis.

4.4.1 Comparative analysis of proposed methods with greedy approach

The proposed method is providing the solution for predicting breast cancer. The Greedy optimization with E-C4.5 algorithm and self-organizing maps with E-C4.5 algorithm are

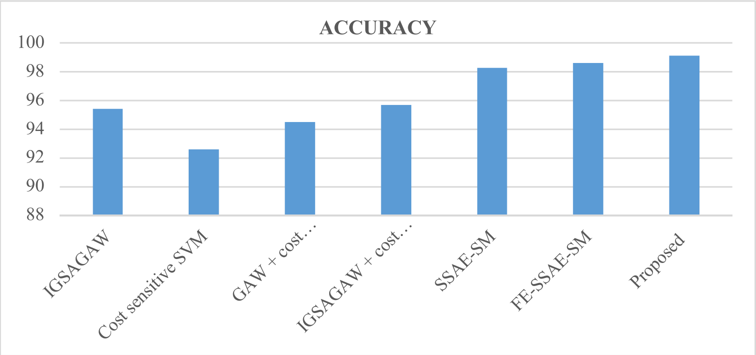

Fig. 8 Comparative analysis of the proposed and existing methods [14] with respect to accuracy

Table 3 Performance analysis of the proposed and existing method [10]

Model	Accuracy	Precision	Recall	F1-score
KNN	95.8	93.5	93.5	93
L_G	95.8	96.5	95	95.5
DT	95.8	96.5	95	95.5
RF	97.2	97	97.5	97
SVM	97.2	97.5	97	97
DL_ANN	98.24	98	98	98
Proposed	99.12	98.5	98.5	98.5

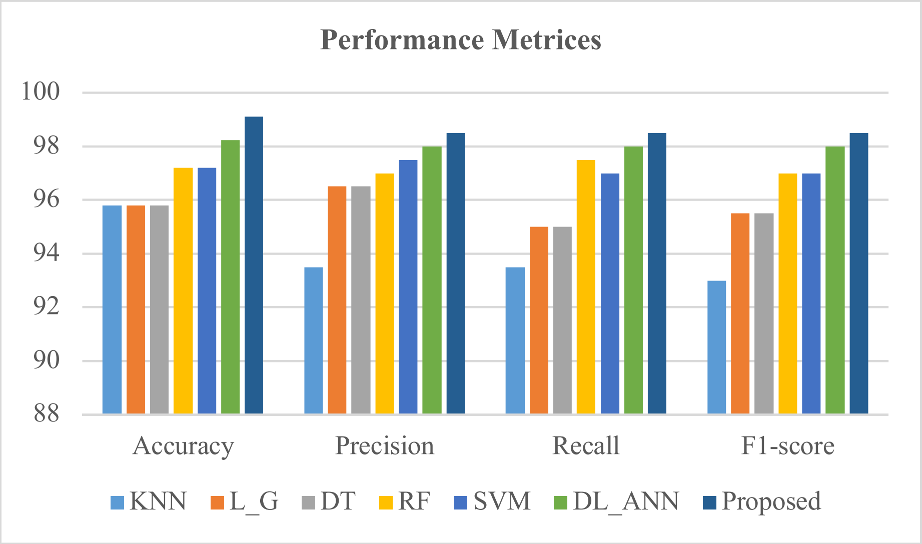

Fig. 9 Comparative analysis of the proposed and existing methods [10]

two methods to predict breast cancer. Here, over all computational is compared with respect to these two methods.

The Table 4 is showing the result of accuracy, precision and F1 score for Greedy approach and proposed method. The comparative analysis shows that the accuracy of Phase 2- self-organizing maps with E-C4.5 algorithm is higher than Greedy optimization with E-C4.5 algorithm with 99.12%.

5 Conclusion

The study proposed methods based on Deep Learning (DL) to predict Breast Cancer (BC). Various processes have been undertaken that included Dimensionality reduction, clustering and prediction. Stochastic Self-Organizing Map (SOM) and Enlarge C4.5 (E-C4.5) algorithm is proposed where Stochastic SOM has been used to cluster the similar data into groups and then predict it by the E-C4.5. The proposed method finds the Malignant (M) and Benign (B) BC counts. The misinterpretation rate of the proposed method is assessed through a confusion matrix and is found that there existed only a few misinterpretations, thereby proving the proposed system's efficacy. A comparative analysis was undertaken to find the efficacy of the proposed method. Random Forest (RF), IGSAGAW, Cost Sensitive Support Vector Machine,

Table 4 Comparative analysis of proposed method and greedy optimization approach

Algorithm	Accuracy	Precision	F1-Score
Greedy optimization with E-C4.5 algorithm	98.06	99	98
Self-organizing maps with E-C4.5 algorithm	99.12	98.5	98.5

eXtreme Gradient Boosting (XGBoost), GAW + cost sensitive SVM, Stacked Sparse AutoEncoders-SoftMax Regression (SSAE-SM), Feature Ensemble SSAE-SM, IGSAGAW + cost sensitive SVM, K-Nearest Neighbour, DT, SVM, Logistic Regression (L_G), and DL_ANN were the numerous existing methodologies taken into account for the comparative analysis. The results showed the efficiency of the proposed methods than the conventional methods with respect to the four performance metrics namely accuracy, recall, F1-score and precision. This study can help the physicians to diagnose BC fast and provide effective treatment to the patients accordingly.

With the integration of advanced techniques, breast cancer prediction will be more helpful in diagnosing the cancer. Hence, the proposed self-organizing maps with E-C4.5 algorithm system is efficient than the other proposed and existing system which is confirmed through the analytical results. This study can help the physicians for early predcition of breast cancer fastly and efficently for treating patients. Thus, this study helps to reduce false diagnosis of humans which might happen due to fatigue.

6 Limitations

The proposed system is only used for predicting breast cancer, it can’t be used commonly for predicting any types of cancer. The proposed self-organizing map is also used for reducing the amount of data, generalization and efficiently compressing information for transmission, speeding up extrapolation and non-linear interpolation. Moreover, this technique has provided the efficient way for predicting the breast cancer but even more accuracy rate can be achieved by enhancing the machine learning algorithm.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Aavula R, Bhramaramba R (2021) Towards a framework for breast cancer prognosis: risk assessment. In: ICCCE 2020. Springer, pp 1517–1533
2. Abdulkareem SA, Abdulkareem ZO (2021) An evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique. International Journal of Sciences: Basic and Applied Research (IJSBAR) 55(2):67–80
3. Arefan D, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S (2020) Deep learning modeling using normal mammograms for predicting breast cancer risk. Med Phys 47:110–118
4. Baloni P, Dinalankara W, Earls JC, Knijnenburg TA, Geman D, Marchionni L et al (2021) Identifying personalized metabolic signatures in breast cancer. Metabolites 11:20
5. Begum A, Kumar R (2021) Review of chronic inflammation and long term effects on health using machine learning algorithms. Int J Comput Sci Eng 9(6):70–76. https://doi.org/10.26438/ijcse/v9i6.7076
6. Benhammou Y, Achchab B, Herrera F, Tabik S (2020) BreakHis based breast cancer automatic diagnosis using deep learning: taxonomy, survey and insights. Neurocomputing 375:9–24
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K et al (2019) A guide to deep learning in healthcare. Nat Med 25:24–29

8. Ghiasi MM, Zendehboudi S (2021) Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med 128:104089
9. Ghosh SK, Mitra A, Ghosh A (2021) A novel intuitionistic fuzzy soft set entrenched mammogram segmentation under multigranulation approximation for breast cancer detection in early stages. Expert Syst Appl 169:114329
10. Gupta P, Garg S (2020) Breast cancer prediction using varying parameters of machine learning models. Procedia Comput Sci 171:593–601
11. Hakkoum H, Idri A, Abnane I (2021) Assessing and comparing interpretability techniques for artificial neural networks breast cancer classication. Computer methods in biomechanics and biomedical engineering: imaging & visualization, pp 1–13
12. Herent P, Schmauch B, Jehanno P, Dehaene O, Saillard C, Balleyguier C et al (2019) Detection and characterization of MRI breast lesions using deep learning. Diagn Interv Imaging 100:219–225
13. Jaiswal A, Kumar R (2020) Review of machine learning algorithms in cancer prognosis and prediction. J All Res Educ Sci Methods 8(6):146–156
14. Kadam VJ, Jadhav SM, Vijayakumar K (2019) Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. J Med Syst 43:1–11
15. Karthik S, Perumal RS, Mouli PC (2018) Breast cancer classification using deep neural networks. In: Knowledge computing and its applications. Springer, pp 227–241
16. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJC (2019) A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognit Lett 125:1–6
17. Kumar V, Mishra BK, Mazzara M, Thanh DN, Verma A (2020) Prediction of malignant and benign breast cancer: a data mining approach in healthcare applications. In: Advances in data science and management. Springer, pp 435–442
18. Kumari M, Singh V (2018) Breast cancer prediction system. Procedia Comput Sci 132:371–376
19. Meenalochini G, Ramkumar S (2021) Survey of machine learning algorithms for breast cancer detection using mammogram images. Mater Today: Proc 37:2738–2743
20. Men K, Zhang T, Chen X, Chen B, Tang Y, Wang S et al (2018) Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Physica Med 50:13–19
21. Nawaz M, Sewissy AA, Soliman THA (2018) Multi-class breast cancer classification using deep learning convolutional neural network. Int J Adv Comput Sci Appl 9:316–332
22. Ravi V, Alazab M, Srinivasan S, Arunachalam A, Soman PK (2021) Adversarial defense: DGA-based botnets and DNS homographs detection through integrated deep learning. IEEE Trans Eng Manage:1–18. https://doi.org/10.1109/TEM.2021.3059664
23. Saha M, Chakraborty C (2018) Her2net: a deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Trans Image Process 27:2189–2200
24. Shaikh K, Krishnan S, Thanki R (2020) Artificial intelligence in breast cancer early detection and diagnosis. Springer, New York
25. Shukla N, Hagenbuchner M, Win KT, Yang J (2018) Breast cancer data analysis for survivability studies and prediction. Comput Methods Programs Biomed 155:199–208
26. Singh A, Dwivedi RK, Kumar R (2021) A survey of lung cancer detection using machine learning techniques for improving classification performance. World J Eng Res Technol 7(4):149–161
27. Sun D, Wang M, Li A (2018) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM Trans Comput Biol Bioinf 16:841–850
28. Wolberg WH, Street WN, Mangasarian OL Breast Cancer Wisconsin (Diagnostic) Data Set creators. University of Wisconsin, Clinical Sciences Center, Madison. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
29. Wu J, Hicks C (2021) Breast cancer type classification using machine learning. J Pers Med 11:61
30. Xie J, Wu Z, Xia Q, Ding L, Fujita H (2020) The differential feature detection and the clustering analysis to breast cancers. In: International conference on industrial, engineering and other applications of applied intelligent systems, pp 457–469
31. Zhang D, Zou L, Zhou X, He F (2018) Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. IEEE Access 6:28936–28944
32. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J et al (2019) Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Informatics 132:103985
33. Zhang Y-D, Satapathy SC, Guttery DS, Górriz JM, Wang S-H (2021) Improved breast cancer classification through combining graph convolutional network and convolutional neural network. Inf Process Manag 58:102439

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.