MACHINE INTELLIGENCE RESEARCH
ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

PROPOSED KNNXGBOOST CLASSIFICATION MODEL FOR BREAST CANCER PREDICTION USING BIG DATA ANALYSIS

[1]Bharathidasan.G and [2*]Dr.A.S.Arunachalam

[1]Research Scholar, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamilnadu, India. Email:bharathidasanastro@gmail.com

[2]Associate Professor, Department of Computer Applications, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamilnadu, India. Email: arunachalam1976@gmail.com

*Corresponding Author

Abstract

The breast cancer identification at early stage remains a difficult task for radiologist and researchers who deal with breast cancer identification process. The big data and image processing coupled together makes a remarkable attempt to solve addressed issues in identification process. The procedure followed in big data mining is much related to the process followed in data mining but differs in handling the huge voluminous confidential medical data. The process followed in the early stage of the research is to collect the breast cancer infected data from WBC, were the preprocessing and feature extraction process is carried out for improving the quality of the collected data as well as to increase the efficiency of the featured breast cancer data. The proposed KNN-XGBOOST classification algorithm is implemented further in segmentation and classification process. The procedure helps in classifying the two different classes segmented in feature identification process using proposed PCA-K-mean clustering technique earlier. The severity levels in breast cancer can be easily identified using the proposed KNN-XGBOOST classification algorithm with minimum time. The proposed classification technique is tested with similar classification techniques for finding the accuracy of identification process. The proposed KNN-XGBOOST performance is better than the other existing classification techniques with best time taken for implantation.

1. Introduction

Computer vision based technology in predicting the presence of cancer plays a major role in identifying the breast cancer in human body. The latest technologies of big data analyses coupled with clinical cancer cell predictions are very useful in improving chance of early detection and increases the chance of remedial actions on breast cancer. Most of the research work carried out the field of breast cancer perdition concentrates on finding the presence of cancer cells on later stage or intermediate stage. The process followed in existing research work are not limited in using computerized technologies rather using big data technologies for better results.

The used data in the research work are huge in number and very confidential healthcare record. Though the data should be maintained with much more care and should be maintained with high security. The Big data is one of the computerized fields were analyzing and visualization of collected data from huge data are maintained with much more care. The processing of the collected data can be much more effective while comparing with other computer vision-based

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

approaches.

With more than 1 in 10 new cancer diagnoses each year, breast cancer is the most common cancer in women. It is the second most frequent cancer-related death among women worldwide. The breast's milk-producing glands are located in front of the chest wall anatomically. They are supported by ligaments that connect the breast to the chest wall and lie on the pectoralis major muscle. The breast is made up of 15–20 lobes that are organised in a circle. The size and shape of the breasts are determined by the fat covering the lobes. Each lobe is made up of lobules that contain the glands that produce milk when hormones are stimulated. Breast cancer develops silently at all times.

The majority of people learn they have their disease during routine screenings. Others might exhibit a breast lump that was discovered by accident, a change in the size or contour of the breasts, or nipple discharge. Mastalgia, however, is a frequent condition. Breast cancer diagnosis requires a physical examination, imaging, particularly mammography, and tissue biopsy[34]. With earlier diagnosis, the survival rate increases. Poor prognosis and distant metastasis are caused by the tumor's propensity to spread lymphatically and hematologically. This clarifies and highlights the significance of breast cancer screening initiatives.

Although the a etiology of breast cancer is complex and still not fully understood, there are several established risk factors. 10% of breast cancers are caused by genetic abnormalities and the rest of the percentage are mostly unidentified due to variations in symptoms. The identification process and prediction procedure followed through clinical procedure takes time and creates many subsequent problem in patient’s health.

Big data and machine learning algorithms are very useful in reducing the barriers of accuracy in earlier identification process and reduces the prediction time drastically. The initial stages followed in big data analyzing process is much like that of the data mining process and only varies in using a huge amount of data in prediction procedure. The difference of the bigdata analyzing stage is seen in the pattern identification stage and rue generation stage.

The pre-processing stage removes the irrelevancies present in the collected data and feature extraction process or selection process is mostly considered as optional one, but this research work uses the feature extraction process for selecting the necessary features wanted in prediction process. The stage next to selection process is to process the collected data into the proposed algorithmic techniques.

Breast cancer datasets collected from WBC that are organised and semi-structured to use for evaluating process. With the aid of a clustering-based technique and principal component analysis (PCA), the dimensional data present in the gathered breast cancer dataset are decreased[29]. The outcomes of the dimensionality reduction procedure are regarded as a novel feature used in the analysis phases. In the testing step, proper training is provided for the new features obtained through dimensionality reductions and sent for creating models in the classification process.

The proposed KNNXGBOOST Classification algorithm is a combination of KNN mean based clustering algorithm and Xgboost classification algorithm, which is very useful in classification stage. The proposed algorithm is tested with various existing algorithm for accuracy and time taken for implementation. The proposed algorithm performance is best compared with other existing algorithms such as Decision Tree, Gradient Boosting, Gaussian Naïve Bayes, KNN, Logistic Regression, Random Forest, SVM Linear and SVM RBF.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

2. Literature Review

Extremely large volumes of data have been created experimentally as a result of the computer simulations. This led to the extensive development of large data processing techniques and technologies as well as shifts in scientific research paradigms, including data-intensive science. Data-intensive scientific discovery (DISD), often known as Big Data issues, is the birth of a new paradigm in science [5]. The adjustment entails new study being conducted and knowledge being discovered through data analysis. During this procedure, significant correlations are looked for, and cutting-edge knowledge discovery techniques are used. Knowledge discovery is based on "data-intensive decision making," which is enabled by the new paradigm's methodologies and technological advancements [6].

Data collection, integration and selection, analysis, and data-intensive decision making are all included in the DISD paradigm[30]. There are also challenges associated with the processing technologies, including data accumulation, storage, searching, sharing, analysis, and visualisation, as well as parallel, distributed, and in memory processing. For the development of personalised medicine, human genetic variety, rare and common mutations connected to the sensitivity of human disease, and genetic diversity are crucial [7]. In accordance with each patient's unique characteristics and, in particular, based on genetic analysis of each individual patient, precision medicine provides appropriate and ideal illness diagnostics, medical decisions, treatments, and therapies [8, 9].

With the development of precision medicine, the workload and complexity of physicians' job in the diagnosis of breast cancer have recently increased. The study of cancer has undergone constant evolution [10]. One of the most dangerous and prevalent tumours that primarily affects women is breast cancer. Up to 10% of all breast cancers have mutations in the BRCA1 and BRCA2 genes, which are responsible for the majority of inherited instances of breast cancer [11]. The likelihood of survival rises with early diagnosis. Consequently, a technique for accurately and consistently diagnosing breast cancer is required. In order to analyse breast cancer data, a variety of data mining and machine learning techniques are available, including image processing and retrieval[31]. One of the most significant and fundamental tasks in machine learning and data mining is classification.

Breast cancer develops when cells start to proliferate uncontrollably. These cells typically develop into tumours, which are frequently detectable on an x-ray or felt as lumps. The tumour is malignant (cáncerous) if the cells can spread (metastasize) to distant areas of the body or grow into (invade) surrounding tissues. Breast cancer mostly affects women, but it can also affect men. Breast cancers can begin in several parts of the breast. Most breast cancers (called ductal cancers) start in the milk-transporting ducts that go to the nipple[35]. Some begin in the breast milk-producing glands (common cancers) [6].

The initial stage is typically when oncology doctors attempt to determine whether the cancer has spread beyond the area where it first appeared. According to these traits, there are two main types of breast cancer: in situ (which hasn't spread) and invasive (which has invaded the surrounding breast tissue).There are several different forms of invasive breast cancer, including adenoid cystic cancer, low-grade adenosquamous cancer, medullary cancer, mucinous cancer, pápillary cancer, and tubular cancer. Less frequent forms of breast cancer include inflammatory breast cancer, Pattee disease of the nipple, Phyllodes tumour, and angiocarcinoma [7].

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

It's crucial to recognises that the majority of breast lumps are benign and not cancerous. Although benign breast tumours are common growths, they do not spread outside of the breast and do not pose a threat to life[32]. However, certain benign breast lumps can raise a woman's risk of developing breast cancer. Any breast lump or change should be examined by a health care provider to rule out cancer and assess whether it may increase the chance of developing the disease in the future. Cancer of the breast can spread through the lymphatic system.

The body's lymphatic system consists of lymph nodes, lymphatic veins, and lymph fluid. Small, bean-shaped collections of immune cells known as lymph nodes are connected by lymphatic (or lymphaticatic) veins. Similar to small veins, lymph vessels carry lymph—a clear fluid—instead of blood out from the breast. Immune system cells, tissue fluid, and waste products are all found in lymph. Breast cancer cells have the ability to enter lymph veins and start growing in lymph nodes. There is a higher chance that cancer cells may have spread (metastized) to other areas in your body if they have reached your lymph nodes[36]. The likelihood that breast cancer cells will also be found in other organs increases with the number of lymph nodes that contain them. Because of this, discovering cancer in one or more lymph nodes frequently has an impact on the treatment strategy.

In order to determine whether the cancer has spread there, one or more lymph nodes will typically need to be removed during surgery.But not all women with cancerous cells in their lymph nodes develop metastases, and some women may not even have cancerous cells in their lymph nodes at first [8]. Early detection is crucial for human life, especially for the aggressive breast cancer forms. According to statistics provided by the World Cancer Research Fund International (WCRFI), around 1.7 million new cases of breast cancer were diagnosed in 2012. It is the second-most prevalent cancer worldwide. About 12% of all new cancer cases and 25% of all cancers in women are represented by this. Many studies have been done to categorise breast cancer data using data mining and machine learning on various medical datasets [12, 13, 14].

Many of them exhibit high categorization precision. Building precise and computationally effective classifiers for precision medicine, specifically for the case study of breast cancer, is a significant challenge in the fields of data mining and machine learning. From the foregoing, it can be inferred that the issue of tailoring a patient's treatment for breast cancer is extremely complex, necessitating extensive examination and analysis on the part of the physician of various types of data, including genetic characteristics, the patient's health, health status, and environmental factors.[26, 27, 28] When it comes to data analysis, big data technology can be quite beneficial for doctors. Scientists can benefit greatly from a complete system for precision medicine that encompasses all stages of data discovery, integration, preprocessing, constructing models, storage, analysis, and result visualisation.

3. Proposed Methodology

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

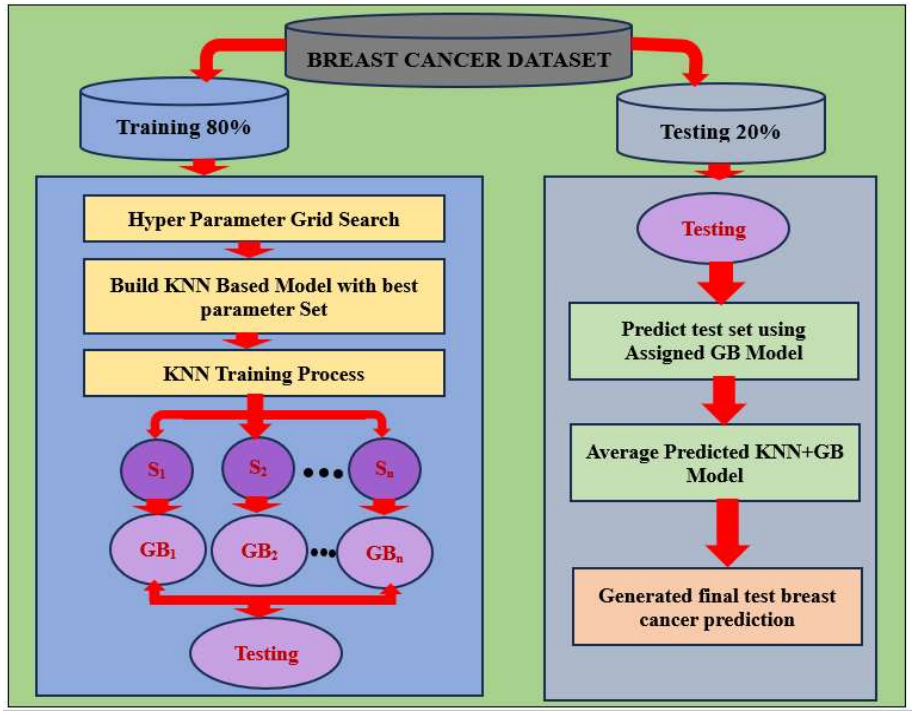

Figure 1. Architectural diagram of proposed KNNXGBOOST model

The primary goal of machine learning approaches is to create a classification model using a dataset that contains labelled classes and some features, such as a dependent binary variable and an independent variable. The training and dataset validation phases make up the majority of the SL and SSL machine algorithms' workflow. The approach modifies the prediction model to reduce error in the output results using the training dataset. The development of the learning algorithm independently because the validation dataset is split from the training dataset. This measure's primary goal is to establish the training algorithm's endpoint in order to stabilise the trained model's accuracy against overfitting.

The most popular machine learning method is SL. Learning a function that matches the input pairs of values to the output is necessary for machine learning. Each input pair corresponds to a labelled value, and the function extracts knowledge from labelled training data. By identifying a pattern in the training data, SL algorithms can create a function that can forecast fresh input pairs or previously unseen observations. The algorithm generalises the function to accurately forecast the hidden.

The procedure carried out in preprocessing feature extraction process clearly segments the breast cancer data from the collected big data database. The rules applied in preprocessing and feature extraction are very useful in enhancing the clarity of the collected breast cancer data from big data[33]. The overall collected breast cancer dataset is divided into two parts for testing and training. Usually, 80% of the overall data is considered for training and 20% is considered for the testing process.

The proposed KNNXGBOOST model takes the necessary steps to use KNN model as well as XGBOOST gradient model for classification process. The Extreme Gradient Boosting (XGBoost) library of gradient boosting algorithms has been specially designed for the prediction

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

process of breast cancer and issues of contemporary data science. The fact that XGBoost is extremely scalable and parallelizable, rapid to execute, and often outperforms other algorithms are some of its main advantages. It also uses a more regularised model formalisation to control over-fitting, which improves performance.

The training phase of the proposed model discusses about the combination of hyper parameter grid search and KNN algorithm in assigning the class weightage. The collected breast cancer dataset are segmented in various iteration is feature extraction process. The processed data is taken into the fixed model consist of finite set of hyper parameters. A parameter that is established before the learning process starts is referred to as a hyperparameter. These adjustable settings have a direct impact on how successfully a model trains. Several instances of machine learning hyperparameters include: Algorithms for machine learning can be trained with the help of hyperparameters.

The next process followed in the proposed KNNXGBOOT model is implementing the KNN machine learning technique in classification procedure. The data points are consisted as i = 1, 2,…,n and the overall set of data are known to be (X_i, C_i), were X denotes the feature values collected form breast cancer dataset and C denotes labeling to the collected classes. The assigned values are given for each value collected from the collected breast cancer dataset. The arrangement of the variables are made according to the nearest neighbor algorithm, which is the most important process in finding out the relationship between each values. The weightage is calculated and given separate naming for different values depending upon the weightage of the values.

The categorized values are shown as S = {S_1, S_2, S_3,…S_n}, were S denotes the each weighted values from the KNN process. The next step is the most important part of the proposed model, which not only improves the seep of the execution, it also increases the accuracy of the prediction process. The Extreme Gradient Boosting (XGBoost) is the most powerful part of the proposed model, which is best suitable for the taking the consolidated data into next level. The collected Set S = {S_1, S_2, S_3,…S_n}, is reassigned with Extreme Gradient Boosting mechanism GBM ={GB_1, GB_2, GB_3,… GB_n}, were GB stans for Gradient Boosted values. The process is followed throughout all the classifies values. The Gradient Boosting phenomena is very useful in categorizing the values according to there weightage. The proposed KNNXGBOOST classification model is explained with the following steps.

Data classification using KNNXGBOOST model

Step 1: Load all Breast cancer Dataset

Step 2: Divide 80% data for Training and 20% for Testing

Step 3: Start the training procedure

Step 4: Initiate the hyper parameter search model

Step 5: Build KNN based model with best parameter set

Step 6: Select the values for K instance

Step 7: Calculate the distance between each data point using Euclidean distance.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

Step 8: Sort data point according to the calculated distance

Step 9: Select the topmost K^{th} row

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

Step 10: Assign data point to the most frequent weighted classes
Step 11: Assign the weighted class to Extreme Gradient Boosting mechanism (GBM)
Step 12: Assign S = $\{S_1, S_2, S_3, \ldots S_n\}$ to GBM = $\{GB_1, GB_2, GB_3, \ldots GB_n\}$ based on weightage
Step 13: Start the testing procedure
Step 14: Predict the test set using assigned GBM
Step 15: Find the average of predicted models
Step 16: Generate the final test predictions
Step 17: End

The proposed KNNXGBOOST model is tested with various existing classification models for finding the accuracy and other constrains for determining the perfect prediction model for breast cancer. The performance of the proposed KNNXGBOOST model plays good role compared with other algorithms.

The collected information after the implementation of proposed PCA-K-Mean algorithm for feature extraction process is taken for the classification stage, were the division of Benign (B) and Malignant (M) are made and compared with existing classification algorithms. The method follows the actual procedure of KNN and taken for the activation function.

The activation function used in the research work plays a major role in triggering the function by setting the parameters for the classification stage manually. This procedure is considered as most advantage stage because most of the existing researches are negative in following manual activation function. Setting manual activation function does not alter the parameter setting throughout the execution and helps in acquiring the accurate prediction.

The procedure followed after the implementation of the KNN procedure is basically followed with XG – BOOSTING algorithmic technique, were the collected weighted instants are arranged hieratical order in order to enhance the quality of the classification procedure. The XG BOOSTING technique improves the weightage of the arranged instances with unique representation. The proposed KNN-XG-BOOSTING classification algorithm in compared with various existing algorithms for testing the performance and efficiency of the proposed algorithm. The time consuming and accuracy are other scenarios considered for measuring the efficiency of the proposed algorithm.

4. Results and Discussion

The proposed KNNXGBOOST Classification algorithm is a combination of KNN mean based clustering algorithm and Xgboost classification algorithm, which is very useful in classification stage. The proposed algorithm is tested with various existing algorithm for accuracy and time taken for implementation. The proposed algorithm performance is best compared with other existing algorithms such as Decision Tree, Gradient Boosting, Gaussian Naive Bayes, KNN, Logistic Regression, Random Forest, SVM Linear and SVM RBF.

4.1. Data Description

The breast cancer data are collected from WBC with 8670 record set and 31 different attributes. The list of collected attributes from the WBC dataset are listed in the table 1. The diagnosis of breast tissues (M = malignant, B = benign) present in the collected dataset are 63% for M and 37% for B.

Table 1. Attribute collected from WBC

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

S. No	Selected Attributes
1.	ID
2.	radius_mean
3.	texture_mean
4.	perimeter_mean
5.	area_mean
6.	smoothness_mean
7.	compactness_mean
8.	concavity_mean
9.	concave
10.	points_mean
11.	symmetry_mean
12.	fractal_dimension_mean
13.	radius_se
14.	texture_se
15.	perimeter_se
16.	area_sesmoothness_se
17.	compactness_se
18.	concavity_se
19.	concave points_se
20.	symmetry_se
21.	fractal_dimension_se
22.	radius_worst
23.	texture_worst
24.	perimeter_worst
25.	area_worst
26.	smoothness_worst
27.	compactness_worst
28.	concavity_worst
29.	concave points_worst
30.	symmetry_worst
31.	fractal_dimension_worst

The sample dataset from collected data are shown in table 2, which also clearly shows the differences in attributes.

Table 2: Sample WBC data

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471
842517	M	20.57	7.77	132.9	1326	0.08474	0.07864	0.0869	0.07017
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279
84348301	M	11.42	20.38	7.58	86.1	0.1425	0.2839	0.2414	0.1052
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076

The table 2 shows the sample of collected breast cancer data, which is divided based in the IDs. The diagnosis of breast tissues malignant and benign is represented with M and B respectively. The radius means of calculating the distance from the center point are also shown. Various mean such as texture mean, perimeter mean, area mean, smoothness means, compactness means, concavity mean and concave point mean are also shown in the table.

The proposed KNN-XGBOOST classification algorithm is used for classification stage, were the diagnosis of M and B are classified accordingly. The proposed algorithm is tested with various existing algorithms for testing performance and accuracy. Multiple measurement like Sensitivity, specificity, recall, precision, F1 score which improves the prediction accuracy are used for measuring the efficiency of the proposed algorithm.

Table 3: Accuracy of the proposed KNN-XGBOOST Classification model

Measurement	Classification algorithms accuracy								
	Decision Tree	Gradient Boosting	Gaussian Naïve Bayes	KNN	Logistic Regression	Random Forest	SVM Linear	SVM RBF	KNNXgboost
Sensitivity	0.90	0.84	0.95	0.95	0.98	0.92	0.97	0.89	0.99
Specificity	0.98	0.99	0.98	0.94	0.99	1.00	0.57	0.93	1.00
Pos Pred Value	0.97	0.98	0.96	0.91	0.98	1.00	0.57	0.88	0.99
Neg Pred Value	0.95	0.91	0.97	0.97	0.99	0.96	0.97	0.93	1.00
Precision	0.97	0.98	0.97	0.91	0.98	1.00	0.57	0.88	0.99
Recall	0.90	0.84	0.95	0.95	0.98	0.92	0.97	0.89	0.99
F1	0.93	0.91	0.96	0.93	0.98	0.96	0.72	0.88	0.99
Prevalence	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37
Detection Rate	0.34	0.31	0.35	0.35	0.36	0.34	0.36	0.33	0.38

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

Detection Prevalence	0.35	0.32	0.36	0.39	0.37	0.34	0.63	0.38	0.38
Balanced Accuracy	0.94	0.92	0.97	0.95	0.99	0.96	0.77	0.91	1.00

Table 3 describes the efficiency of the proposed KNN-XG-BOOST algorithm in concerning the Sensitivity, Specificity, Pos Pred Value, Neg Pred Value, Precision, Recall, F1, Prevalence, Detection Rate, Detection Prevalence and Balanced Accuracy. The feature extraction stage explains that the Logical Regression plays a best role in efferent way of the classification, but in the final stage of the research work with the usage of KNN-XG-Boosting algorithm. The proposed KNN-XG-BOOSTING algorithm is slightly improving the accuracy and time than that of the other proposed PCA-K-Mean algorithm. The proposed algorithms are tested with other existing algorithms with various measurements and found that the proposed KNN-XG-BOOSTING performance is better then other existing algorithms.

Table 4: Time consumed during Training and testing set

Modelling Method	**Training (Sec)**	**Testing (Sec)**
Decision Tree	0.1590	0.0498
Gradient Boosting	1.3080	0.0450
Gaussian Naïve Bayes	0.0551	0.0479
KNN	0.0010	0.0100
Logistic Regression	0.1890	0.0798
Random Forest	0.1590	0.0898
SVM Linear	0.1590	0.0998
SVM RBF	0.1590	0.1598
KNNXGBOOST	0.0002	0.0080

The execution time for the proposed KNNXGBOOST algorithm is also compared with other classification algorithms as in table 4. The proposed execution time is much lesser than the other existing algorithms. The proposed KNNXGBOOST model is better in executing the testing time and training time also.

Table 5: comparison Area under the curve (AUC of ROC curves

Modelling Method	**AUC of ROC Curve Supervised**	**AUC of ROC Curve Semi-Supervised**
Decision Tree	0.90	0.91
Gradient Boosting	0.95	0.90
Gaussian Naïve Bayes	0.96	0.99
KNN	0.96	0.97
Logistic Regression	0.98	0.92
Random Forest	0.99	0.98

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

SVM Linear	0.98	0.97
SVM RBF	0.97	0.97
KNNXGBOOST	0.99	0.93

The comparison of AUC and ROC also better in proposed KNNXGBOOST model and reasonably other existing algorithms are also shows best performance in comparing the AUC and ROC as in Table 5. A logistic regression model repeatedly can be analyzed with various classification criteria to generate the points on a ROC curve, but this would be inefficient. Fortunately, the efficient sorting-based method known as AUC can give perfect information. An overall measure of performance across all potential classification criteria is provided by AUC. AUC can be seen as the likelihood that the model values a randomly chosen positive example higher than a randomly chosen negative instance.

5. Conclusions

The classified malignant tumour data and benign tumour data from the collected breast cancer record set is taken for the feature identification process. The process is carried out with K-mean clustering concept, where necessary arrangements are made for securitizing the necessary feature from 31 breast cancer attributes. The final classification stage of the research work also implements with KNN-XG-BOOSTING algorithm for testing the efficiency in predicting breast cancer in WBC dataset. The comparisons are carried out with various existing algorithms and proposed algorithms KNN-XG-BOOSTING algorithm. The KNN-XG-BOOSTING algorithm is used for final classification procedure, which plays a major role in analyzing the breast cancer effected data accurately and with minimum time duration. The results and discussion suggest that the proposed KNN-XG-BOOSTING model performance is better in classification process, than other classification models. The proposed KNN-XG-BOOSTING classification algorithm performance is slightly more that the Logical Regression used in classification stage.

Reference

[1]. AbhijitRaorane&R.V.Kulkarni, “Data Mining Techniques: A Source for Customer Behavior Analysis”, 2011.

[2]. A. Yates, NATIONAL ACADEMY OF SCIENCES, 1997. WASHINGTON DC.

[3]. C.E. DeSantis, J. Ma, M.M. Gaudet, et al., Breast cancer statistics, CA A Cancer J. Clin. 69 (6) (2019) 438–451.

[4]. N. Harbeck, F. Penault-Llorca, J. Cortes, et al., Breast cancer, Nat Rev Dis Primers 5 (1) (2019) 66.

[5]. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[6]. M. Shi, B. Zhang, Semi-supervised learning improves gene expression-based prediction of cancer recurrence, Bioinformatics 27 (21) (2011) 3017–3023.

[7]. S. Becker, A historic and scientific review of breast cancer: the next global healthcare challenge, Int. J. Gynaecol. Obstet. 131 (1) (2015) S36–S39.

[8]. A.R. Padhani, G. Liu, D.M. Koh, et al., Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations, Neoplasia 11 (2) (2009) 102–125.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

[9]. T. Choi, S. Park, J. Oh, Realization method for No ActiveX using emscripten, Korean Society For Internet Information 15 (2014) 49–50.
[10]. D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.
[11]. K.U. Rani, Parallel approach for diagnosis of breast cancer using neural network technique, Int. J. Comput. Appl. 10 (3) (2010) 1–5.
[12]. A.S. Sarvestani, A. Safavi, N. Parandeh, M. Salehi, Predicting Breast Cancer Survivability Using Data Mining Techniques, IEEE, 2010.
[13]. Bharathi.A and A.S.Arunachalam, Feature Extraction for Identifying Alzheimer's Disease Using Deep Learning, NeuroQuantology, Volume20,Issue10, 2022.
[14]. Bharathidasan.G and A.S.Arunachalam, Pre Processing for Early Detection of Breast Cancer using Machine Learning, NeuroQuantology, Volume20,Issue10, 2022.
[15]. L.H. Sobin, M.K. Gospodarowicz, C. Wittekind, TNM Classification of Malignant Tumours, John Wiley & Sons, 2011.
[16]. C. Sotiriou, S.Y. Neo, L.M. McShane, et al., Breast cancer classification and prognosis based on gene expression profiles from a population-based study, Proc. Natl. Acad. Sci. U. S. A. 100 (18) (2003) 10393–10398.
[17]. C. Shravya, K. Pravalika, S. Subhani, Prediction of breast cancer using supervised machine learning techniques, Int. J. Innovative Technol. Explor. Eng. 8 (6) (2019) 1106–1110.
[18]. N. Nikolaou, H. Reeve, G. Brown, Margin Maximization as Lossless Maximal Compression, 2020, 200110318.
[19]. C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, 2017.
[20]. D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Appl. Soft Comput. (2019), 105524.
[21]. Y.C.P. Reddy, P. Viswanath, B.E. Reddy, Semi-supervised learning: a brief review, Int. J. Eng. Technol. 7 (1.8) (2018) 81.
[22]. Dua D,Graff C. 2019. R. Agha, A. Abdall-Razak, E. Crossley, et al., STROCSS 2019 Guideline: Strengthening the reporting of cohort studies in surgery, Int. J. Surg. 72 (2019) 156–165.
[23]. Zhang, J. Xu, X. Hu, et al., Diagnostic method of diabetes based on support vector machine and tongue images, BioMed Res. Int. 2017 (2017).
[24]. J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Canc. Inf. 2 (2006), 117693510600200030.
[25]. S.H.S.A. Ubaidillah, R. Sallehuddin, N.H. Mustaffa, Classification of Liver Cancer Using Artificial Neural Network and Support Vector Machine, 2014.
[26]. Zsigmond, T., & Szeberényi, A. (2023). Környezettudatos fogyasztói magatartás a felvidéki fogyasztók körében (Green consumer behaviour among consumers in the Uplands). In: Kovács László, Szőke Viktória (ed.) A zöld üzleti gondolkodás és a zöld marketing lehetőségei és kihívásai. Szombathely: Savaria University Press, pp. 167-183.
[27]. Rokicki, T.; Koszela, G.; Ochnio L.; Perkowska A.; Bórawski, P.; Bełdycka-Bórawska A.; Gradziuk B.; Gradziuk P.; Siedlecka A.; Szeberényi A.; Dzikuć M. Changes

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

in the production of energy from renewable sources in the countries of Central and Eastern Europe. Frontiers in Energy Research 2022, 10, 993547.
[28]. Szeberényi, A.; Rokicki, T.; Papp-Váry, Á. Examining the Relationship between Renewable Energy and Environmental Awareness. Energies 2022, 15(19), 7082.
[29]. Ganvir, V. Y., Ganvir, H. V., &Gedam, R. S. (2022). Effect of lanthanum oxide addition on physical, electrical and dielectric properties in lithium borosilicate glasses. Ferroelectrics, 587(1), 127-138.
[30]. Ganvir, H. V., Ganvir, V. Y., &Gedam, R. S. (2022). Investigation of structural and electrical properties of nickel chloride doped pyrrole aniline copolymer. Materials Today: Proceedings, 49, 1827-1832.
[31]. Ganvir, V. Y., Ganvir, H. V., &Gedam, R. S. (2022). Physical and optical study of Nd2O3 doped sodium borosilicate glasses. Materials Today: Proceedings, 51, 1201-1205.
[32]. Ganvir, V. Y., Ganvir, H. V., &Gedam, R. S. (2019). Effect of Dy2O3 on electrical conductivity, dielectric properties and physical properties in lithium borosilicate glasses. Integrated Ferroelectrics, 203(1), 1-11.
[33]. Wasnik, H. R., Kelkar, D. S., &Ganvir, V. Y. (2015). Yield analysis of copolymers: effect of temperature, feed ratio and initiator concentration on the copolymerization. Journal of Polymer Engineering, 35(2), 99-103.
[34]. Durga Bhavani, K., Ferni Ukrit, M. Design of inception with deep convolutional neural network based fall detection and classification model. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-16476-6
[35]. K. Durga Bhavani, Dr. Radhika N. (2020). K-Means Clustering using Nature-Inspired Optimization Algorithms-A Comparative Survey. International Journal of Advanced Science and Technology, 29(6s), 2466-2472.
[36]. K. D. Bhavani and M. F. Ukrit, "Human Fall Detection using Gaussian Mixture Model and Fall Motion Mixture Model," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 1814-1818, doi: 10.1109/ICIRCA57980.2023.10220913.