International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Supervised Machine Learning-Based Classification and Prediction of Breast Cancer

Dr. Sumeet Mathur[1] and Mr. Sandeep Gupta[2]

Submitted: 27/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: There is currently no cure for breast cancer, despite the fact that it is one of the deadliest diseases afflicting women. Each year, the number of fatalities from breast cancer rises dramatically. The predominant form of cancer that results in mortality among women is this particular malignancy, which is of global distribution. To live a long and healthy life, any progress in the detection and treatment of cancer is essential. Therefore, maintaining the treatment aspect and patient survival level requires a high degree of accuracy in cancer prognosis. Detection of breast cancer has been facilitated by ML techniques ever since the inception of AI. This has allowed for an earlier diagnosis, hence improving the prognosis for patients. Researchers are paying close attention to ML methods because of their effectiveness; these approaches may soon have a major influence on the prediction and early diagnosis of breast cancer (BC). This article presents a method for automated BC screening that is based on ML. To facilitate an early detection of BC, numerous classification and prognostic models are developed in this research. These models rely on ML techniques, including gradient boosting and lung branching modelling. To test and train their models, the researchers in this work used an UCI ML Repository's BC Wisconsin (Diagnostic) dataset (BCWD). Additionally, an impact of feature selection, data balance, and data preparation methods used on the input dataset. An aim of this study is to identify a best ML algorithms for predicting and diagnosing BC using metrics including F1-score, accuracy, precision, recall, and confusion matrices. A result show that LGBM performed a best among the classifiers and had the greatest accuracy (98%). Based on the Python programming language and associated libraries, all work is completed in the Jupyter notebook environment.

Keywords: *LGBM, Machine Learning, Breast Cancer, Gradient Boost, Wisconsin Dataset.*

1. Introduction

Among the malignancies taking the lives of women is breast cancer. Different types of breast cancer are distinguished by a tumor's location, which include the ducts, lobules, and middle tissue. But generally speaking, they may be divided into benign and malignant forms [1]. A benign tumour does not spread to the other breast tissues, but malignant tumours do. Although mammography reduces the mortality rate from breast cancer by screening breast tissue, it has significant limitations. In the past, several technologies have been utilised to detect breast cancer. Ultrasound imaging is another method of identification that involves inserting ultrasonic waves into the body. However, it has several drawbacks, such as its inability to recognise tumours smaller than 5 mm in size. Sonography is an additional tool for tumour identification. Mammography is used first, and sonography is used to check for any abnormalities. Thermography of the breasts by use of infrared sensors allows for the imaging of temperature changes inside the breast tissue. High-temperature areas are regarded as tumours. These technologies were all somewhat dangerous and produced findings that were not entirely accurate.

The development of a breast cancer prediction model that incorporates all known risk indicators is an enormous undertaking [2][3]. Present prediction methods may neglect other important aspects in favour of analysing demographic risk factors or mammographic images. Moreover, these models might lead to intrusive sampling using MRI and ultrasound, as well as repeated screenings, as they are accurate enough to identify women at high risk. Patients may have to deal with the emotional and financial strain [4][5]. Factors such as demographics, test results, and mammography are necessary for accurate risk prediction of breast cancer [6]. A more accurate evaluation of the likelihood of breast cancer might be accomplished via the use of multifactorial models that include a multitude of risk indicators [7].

Length, lymph nodes, disease dissemination, and invasiveness/noninvasiveness of most tumours are the factors that influence the degrees of breast cancer. Ranges for breast cancer may also be classified as local, close by, and far away [8]. The majority of BC are limited to a breast, yet they might occur close together. It's localised when the majority of tumours are found in lymph nodes, which are often found under the armpit. The term "remote stage" describes the point at which breast cancer has spread to other organs. Furthermore, the majority of malignancies are explained by TNM, another staging

[1] *University of Waikato NZ - Joint Institute at Zhejiang University Hangzhou, China*
[2] *Techieshubhdeep it solutions pvt. Ltd, Gwalior, M.P. India*
ORCID ID: 0000-0003-0752-2381
Corresponding Author Email: ceo.techies@gmail.com

approach. These variables consider the lymph node count (N), tumour size (T), and metastasis (the degree of cancerous tissue metastasis).

- **Stage 0:** DCIS (ductal carcinoma institutions) and other non-invasive breast cancers may be partially explained by this degree. At this point, there is no sign of cancer cells growing on any breast tissue or infecting nearby healthy tissue.
- **Stage I:** Stage I depicts the invasion of normal surrounding breast tissue by the majority of breast tumours. There are no lymph nodes affected at this stage, and the tumour may develop to a size of two centimetres. It is also possible for stage I breast cancer to have a microscopic invasion. The majority of cancer cells that invade in a microscopic manner first begin to infiltrate the tissue outside the duct or lobule's lining; nevertheless, these invaders typically measure little more than 1 mm.
- **Stage II:** Breast cancer is expanding, but it is still confined to the breast or has only migrated to the surrounding lymph nodes. There are two companies in this stage: Stage 2A and Stage 2B. Whether or not the majority of breast cancers have progressed to the lymph nodes and the size of the tumour influence the difference.
- **Stage III:** The three groups that make up Stage III are IIIA, IIIB, and IIIC. Invasion breast cancers are classified as stage IIIA when they have spread to lymph nodes around the breastbone, are located in auxiliary lymph nodes, or are not associated with any other systems. Most of these tumours do not have a tumour.
- **Stage IV:** Most cases of stage IV invasive breast cancer include the disease spreading to other organs such as the lungs, distant lymph nodes, bones, pores and skin, liver, and brain in addition to the breast and surrounding lymph nodes. According to medical terminology, breast cancer classified as level IV is progressed or "metastatic."

Modern technology is highly developed, with newer models yielding more accurate findings than previous ones. ML is a subset of AI, which enables systems to function better without the need for programming by learning on their own using ML algorithms. Doctors want to differentiate between these tumours using a high-quality diagnostic procedure. Even for experts, tumours are often quite difficult to diagnose. Tumour diagnosis also requires the diagnostic gadget to be automated. Many researchers have been using machine learning approaches to identify cancer early. The most popular methods for detecting breast cancer are ML approaches, which are user-friendly and secure for patients [9].

1.1. Contribution of the Paper

This work aims to generate a forecasting model for early diagnosis of breast cancer employing ML algorithms, which may enhance patient prognosis and survival rates by providing prompt therapeutic treatment. This study's main contribution is as follows:

- To collect a BC Wisconsin (Diagnostic) dataset by an UCI ML repository.
- To perform preprocessing task for eliminating null values and check duplicate value.
- To extract a most relevant and discriminating features appropriate for a collected dataset.
- To apply machine learning technique for implementation of breast cancer detection.
- To confirm a validity of an algorithm and the developed prototypes using a testing set of the dataset.
- To assess a constructed model's performance employing assessment measures like F1-score, recall, accuracy, and precision.

1.2. Structure of the Paper

This research is structured as follows for a parts that follow: **Section 2** review the systementic literature review on the breast cancer detection. **Section 3** decribe the research approach that utilized in this paper. In **Section 4**, they cover an outcomes and disscusion of a research project we had in mind. Our research study's findings and plans for the future form **Section 5.**

2. Related Work

Breast cancer detection studies from the past are featured in this section. Few studies have been started with a technological emphasis for BC forecasting, despite a fact that multiple have been reported for analysis of the disease. Here are a few examples of similar works:

In 2023, Das et al.,[10] Using ML methods like SVM and RF has improved sensitivity and precision. Results from several classifiers are compared, and SVM 97% emerges as the top performer according to sensitivity, accuracy and precision. Both benign and malignant cases have a precision of around 97%, whereas benign cases have a sensitivity of 98% and malignant ones of 95%.

In 2023, R. H. Khan, [11] suggested the most effective model for effectively detecting this epidemic. In this study, they investigated an ability of five ML techniques (XGBoost, NB, RF, DT, and LR) to forecast health-related behaviour in humans. XGBoost outperformed the other

algorithms in terms of F-1 score (99%), sensitivity (98.5%), specificity (97.5%), and accuracy (95.42%). According to our research, XGBoost may be a useful tool for breast cancer prediction.

In 2022, Jamal et al. [12] proposes a model of detecting breast cancer that is based on machine learning. Five different ML algorithms were evaluated. 94.73% accuracy was obtained with logistic regression, 92.98% accuracy with DT, 98.24% accuracy with RF, and 96.49% accuracy with SVM. The best accuracy, 98.24%, was provided by RF.

In 2022, C. Roy, [13] provide a substitute for the conventional diagnosis technique by using a variety of ML methods. ML is a non-invasive technique that has a high accuracy rate for detecting breast cancer. By using the DT, RF, LR, and eXtreme gradient boosting techniques, the proposed method exhibited accuracy rates of 92.98%, 96.49%, 97.36%, and 98.14%, respectively.

In 2022, Anklesaria et al., [14] combine a number of ML algorithms with feature selection or hyperparameters, such as SVM, LR, KNN, DT, RF, ANN, and NB. These models were trained employing a WDBC Dataset. Additionally, they found that by using both SMOTE and Undersampling to balance the dataset, Undersampling produced a superior overall outcome. The study found that the SVM Algorithm, which suited our dataset with an accuracy of 95.8%, was the most successful model, followed by KNN, which had an accuracy of 95.3%.

In 2022, M. P. Behera, [15] used five distinct machine learning techniques on the BC dataset: RF, DT, SVM, KNN, and LSTM. All four classifiers— RF, DT, SVM, and KNN —will be evaluated against a LSTM classifier using a following metrics: recall, accuracy, precision, confusion matrix, and F1 score. Predicting the likelihood of breast cancer utilizing ML is a major focus of this research. A results show that, with 96% accuracy, a LSTM algorithm performs better than the other described methods.

In 2022, H. Sharma, [16] utilised the Wisconsin dataset to give a comparative review of contemporary modern ML approaches that are widely employed in cancer diagnosis, notably breast cancer. Classification ML algorithms including KNN, LR, RF, SVM, NB, XG, and DT have been statistically and comparably tested and compared to find a best one according to accuracy, precision, recall, F1 score, & accuracy percent. Additionally, a ROC curve was used to project these classification techniques. Consequently, this study concludes that whereas SVM yields an accuracy of 96.49%, XGboost achieves 98.24% accuracy.

In 2021, H. Sami, [17] advocated for the use of microwave signals in the prediction of breast lesions. Within the realm of biomedical applications, machine learning has consistently shown its reliability in disease detection. They train and evaluate the SVM method using raw data from backscattered signals. The approach uses a linear and polynomial kernel. SVM using a third-degree polynomial kernel achieved 99.7 percent accuracy, outperforming a most advanced traditional ML binary classification method. As a result, radiologists would be able to employ cancer presence prediction to accurately diagnose tumours in their early stages.

In 2020, V. A. Telsang [18] demonstrate an use of several ML algorithms to forecast breast cancer and evaluate their predictive power, AUC, and other performance metrics. The Wisconsin Dataset of Breast Cancer (WDBC) is being used for the purpose of simulation. After all the data was analysed, the SVM model had an AUC of 99.4 and an accuracy of 96.25 percent. In addition, by adjusting the algorithms' mathematical models, we may improve the accuracy of breast cancer predictions.

In 2018, Khuriwal and Mishra, [19] proposed an adaptive ensemble voting approach for breast cancer diagnoses utilising a Wisconsin Breast Cancer database. After diagnose breast cancer with fewer variables, this study compares and explains how ANN and logistic algorithms operate with ensemble ML techniques. Another ML algorithm demonstrated that the ANN method with the logistic algorithm attained an accuracy of 98.50%.

Table 1. Comparative Analysis of Breast Cancer Detection Utilizing Various Approaches

Ref	Methodology	Dataset	Findings	Limitations & future work
[20]	applied ML algorithms	Wisconsin Breast Cancer Dataset and MIAS Dataset	95%	A characteristics taken into consideration for the analysis of breast cancer and the training data set place a limit on the accuracy that machine learning models can deliver.
[21]	Using traditional classifiers. Neural networks are a very advanced subset of ML models.	Wisconsin breast cancer dataset	95.3%	To improve accuracy, it is advised that the extracted dataset's sample size be increased in subsequent work.
[22]	Naive Bayes, J48 Decision Tree, and the Bagging method are	UC Irvine machine learning repository	74.7%	The Simple Logistic Classifier seems to have the most potential to greatly enhance the traditional

	examples of ML algorithms.			classification techniques used in the research, according to a findings of a ML algorithm testing.
[23]	algorithms for ML and clinical data. SVM are used in both supervised (Relief algorithm) and unsupervised (Autoencoder, PCA algorithms) versions of the suggested technique.	Breast Cancer Wisconsin (Original) WBC dataset	99.91%	The main issues with these existing techniques are their large calculation times and poor accuracy, which may be caused by the data set's inappropriate feature selection. To address these problems, new methods for accurately detecting BC are needed.
[24]	Machine learning methods and algorithms, such as RF, kNN, and NB, are often used in cancer prediction.	Wisconsin Diagnosis Breast Cancer data set	94.7%	The Naïve Bayes method just addresses classification issues, whereas kNN and Random Forest are capable of handling both classification and regression difficulties.
[25]	The five methods of supervised ML are as follows: LR, ANNs, RF, KNN, and SVM.	Wisconsin Breast Cancer dataset	98.57%	One neuron is present in the output layer as the issue was one of binary classification. A five-batch batch size was used to adjust the model over seventy epochs.

Recent research on breast cancer diagnosis utilising machine learning has shown significant progress in utilising different techniques to achieve high accuracy rates. However, there are significant gaps in the studies. To start, ensemble approaches, which employ a variety of models to improve prediction performance, have received minimal attention. Secondly, the necessity for standardization is further emphasised by the fact that direct comparisons are made more difficult due to variations in the usage of assessment measures between research. Concerns over the models' capacity to generalize to other populations are further heightened by the difficulty of cross-dataset validation. Last but not least, new studies can strengthen and enhance breast cancer detection models by using cutting-edge methods like deep learning and transfer learning. If these gaps could be filled, ML applications for early detection of breast cancer may be improved and advanced.

3. Research Methodology

The technique of the proposed task is thoroughly explained in this section. There are many subsections within this section. It describes the short description of dataset and how it will preprocess.

3.1. Problem Statement

Breast cancer ranks second in terms of overall mortality. After analysing breast cancer data, they discovered that India has a relatively low number of specialists for breast cancer diagnosis when compared to other nations, resulting in an increase in diagnostic time. The additional time it takes to detect a disease, such as breast cancer, may be deadly for certain people and raises death rates. After automating the diagnostic procedure, the issue of late detection may be eliminated. One application of this concept is a detection of BC with an use of ML algorithms. This will assist in lowering the death rate, reducing dependency on professionals, and cutting down on the expense and duration of diagnosis.

3.2. Methodology

Even today, many developing nations lack the technology necessary to identify breast cancer, a most common disease among women. Find cases of breast cancer in this research by using ML methods. A research approach divide in many phases and steps shoes in ***figure 1***. First, compile a Breast Cancer Wisconsin (Diagnostic) dataset by an UCI ML repository. Then preprocess an original data for check null values, and check duplicate value. Use the SelectKBest approach to choose the best features during the preprocessing stage. eliminate the impact of various dimensions notions on the model's output; standard scalers are used to standardise the data. Using the oversampling approach, the training and test sets are extracted proportionately based on different categories in order to address the issue of data imbalance. Create subgroups for testing and training from the preprocessed data. A divided dataset ratio is 80:20. Machine learning techniques such as gradient boost and LGBM may be used for the classification. After assess a models using accuracy, f1-score, recall, and precision as performance metrics. Our models' performance will improve when they assess the machine learning models.

3.2.1. Data collection and Preprocessing

The very first and typical phase in this study is data collection. In this study, collect BC Wisconsin (Diagnostic) dataset[1] by UCI ML repository webpage. Data preprocessing is an essential part of ML as it enhances the data's quality, which in turn permits the

[1]https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

extraction of useful insights [26]. The term "data preparation" describes the steps used in ML to get raw data ready for use in building and training models. The key terms of data processing are as follow:

- **Check null value:** It is possible for data to be incomplete, meaning that certain values are missing or null. Therefore, there is a predetermined set of procedures to fill in the gaps and deal with missing data.
- **Check duplicate value:** In order to check whether a record is duplicated or not, I can exploit the duplicated ().

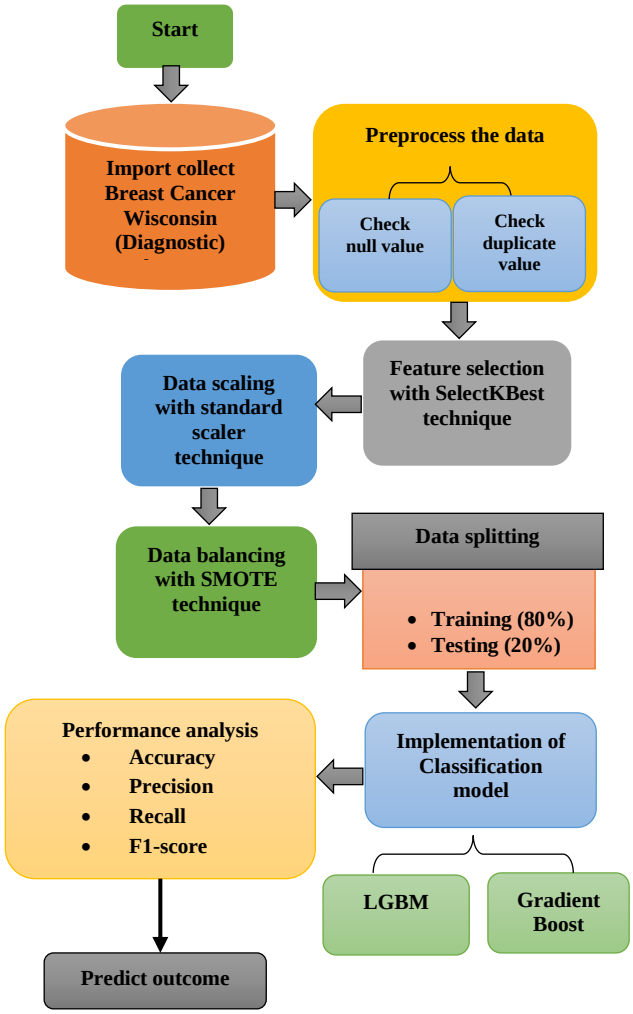

Fig. 1. Proposed flowchart

3.2.2. Feature Selection (SelectKBest technique)

Effective and efficient feature selection has been shown for machine learning tasks. One goal of feature selection is to make models easier to understand and use in data mining. Another goal is to improve data mining performance in areas like prediction accuracy and readability. To create comprehensible data, it also entails being ready to eliminate irrelevant and redundant information [27]. The SelectKBest technique uses the k highest score to choose the features in order to extract them. Both regression and classification data may be used with this approach by varying the parameter. When preparing a big dataset for training, one of the most crucial steps is selecting the optimal features. It assists us in cutting down on training time and removing less significant portions of the data.

3.2.3. Data scaling with Standard scaler

Z-score normalisation is accomplished using the Standard Scaler method. It finds the mean of each value and standardises the characteristics by splitting the outcome by the normal deviation of the attribute. This yields a distribution with unit variance and zero mean. Equation 1 may be used to translate (scale) a value xi into x 0 i, where ¯x is the mean of the x variable.

$$x_j' = \frac{x_i - \bar{x}}{S} \ldots\ldots (1)$$

A sample mean of a property in this instance serves as the translational term, while the standard deviation serves as the scaling factor. This technique may provide a distribution that is very similar for both positive and negative valued variables.

3.2.4. Data balancing with SMOTE

Classification difficulties with imbalanced data might be problematic. A distribution of classes in the training set that is not equal is referred to as unbalanced data. Models with low predictive accuracy might result from this, particularly for a minority class [28] .

After address an issue of class imbalance, our study produces This project seeks to improve accuracy by using the SMOTE and RUS to various datasets. Unlike oversampling by duplication or replacement, SMOTE oversampling approach employs a way of producing arbitrary instances [29].

Algorithm: SMOTE concept

Step 1: let x be a vector representing the sample you want to base your replication on

Step 2: let y be x's nearest neighbor

Step 3: compute d ← y − x

Step 4: let r be a random number ∈ (0, 1)

Step 5: compute the new sample z ← x + (d · r)

3.2.5. Data splitting

Data splitting is a process of distribution of data for the intend of training and testing. In this paper, data split into two sets training and testing. Twenty percent of the data is for testing, whereas eighty percent is for training.

3.2.6. Classification techniques

For model implementation, choose ML approaches for the aim of constructing models. In ML, computers take in data

as input and use statistical analysis to learn new ideas, such as how to categorize and forecast data. For this research, the gradient bost and lightGBM use detect breast cancer.

3.2.6.1. Gradient boost model

A regression method like boosting is called gradient boosting [30]. Gradient boosting aims to minimise an expected value of a certain loss function on a given training dataset $D = \{x_i, y_i\}_1^N$ in order to get an approximation value, F(x), of a function f (x), which links instances x with their corresponding output values y, L(y, F(x)). As an additive estimate of f(x), GB produces the following weighted sum of functions:

$$F_m(X) = F_{m-1}(X) + p_m h_m(X), \ldots\ldots (2)$$

where pm is the mth function's weight, hm. An ensemble's models are these functions. They build the estimate iteratively. First, the following method yields a constant approximation of f (x):

$$F_0(x) = argmin \sum_{i=1}^{N} L(y_i, \alpha). \ldots\ldots (3)$$

The following models must be used in order to minimise.

$$P_m, h_m(x)) = \arg\ min_{p,h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i + ph(x_i)). \ldots. (4)$$

Every hm represents a step in the greedy step gradient descent optimisation for F*. To do so, each model, hm, is trained on a fresh dataset $D = x_i, r_{mi}\}_{i=1}^N$, with pseudoresiduals, rmi, produced as follows:

$$r_{mi} = [\frac{\partial L(y_i, F(x))}{\partial F(x)}]_{F(x)=F_{m-1(x)}} \ldots\ldots (5)$$

where a calculation of Pm is achieved by the resolution of a line search optimisation issue [31]. The Gradient Boosting technique is used in the described approach with the following parameters:

- "max_depth" =1
- "n_estimators" =600)

3.2.6.2. LightGBM

The GBDT (Gradient Boosting Decision Tree) algorithm is implemented by the LightGBM (Light Gradient Boosting Machine) framework [32], which offers distributed support for rapidly processing large amounts of data and supports effective parallel training, lower memory consumption, faster training speeds, and enhanced accuracy. Rather of using a level-wise decision tree development technique that is utilized by other GBDT tools, it employs a leaf-wise algorithm with depth constraints. The strategy divides the leaf that has the biggest split gain among all of the leaves that are currently in use, repeating the cycle. Thus, the benefits of leaf-wise over level-wise are that it may achieve greater accuracy and lower mistakes under the same number of splits. But as Figure 2 below illustrates, it may also lead to overfitting by causing deeper decision trees to form.

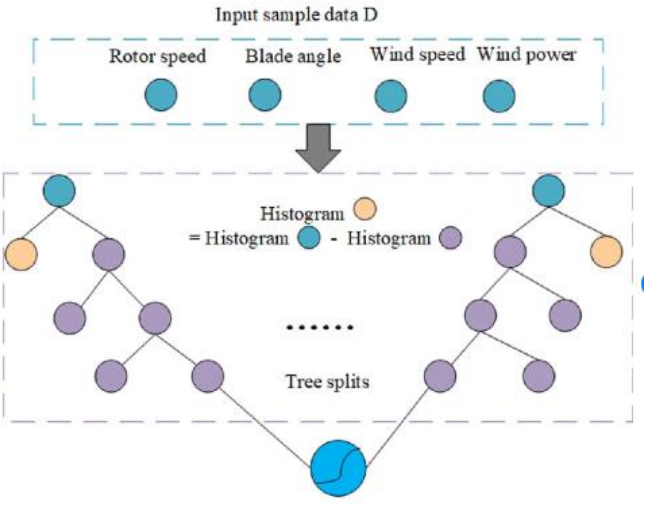

Fig. 2. LightGBM Leaf-wise tree growth

To maximise training time and reduce memory consumption and avoid over-fitting difficulties in LightGBM, one possible technique is to use a histogram-based method in conjunction with the leaf-wise growth strategy of trees that have a maximum depth constraint. Here are other hyperparameters that may be tweaked: "num_leaves," which says how many leaves a tree has; "max_depth," which says how deep a tree can go; and "learning_rate," which is used to equalise a weights of the classes [33]. Finding an appropriate range for this method is necessary to improve the optimisation outcomes. Light Gradient Boosting has the following parameters:

- n_estimators" =600
- max_depth=3

3.2.7. Hyperparameter of proposed models

Hyperparameters are crucial to ML algorithms because they directly regulate how training algorithms behave and greatly impact how well machine learning models function. Numerous methodologies have been devised and effectively used within certain application fields.

Hyperparameter optimisation may be accomplished by using an exhaustive grid search to find an optimal combination of hyperparameters, which improves an accuracy and performance of a model. Hyperparameter optimisation may greatly improve the overall efficacy of machine learning models, despite the fact that this method's exhaustive search technique may make it computationally costly.

3.3. Proposed Algorithm

Proposed Algorithm: breast cancer detection

Step 1: Python and Jupyter Notebook should be installed.

- Import python libraries like matplotlib, accuracy_score, precision, recall, f1-score etc.

Step 2: Data Collection

- Collect a BC Wisconsin (Diagnostic) dataset.
- Dataset collect by UCI ML repository.

Step 3: Data Preprocessing

- Preprocess the data for check null value and check duplicate value.

Step 4: Feature Selecton with SelectKBest method.

Step 5: data scaling with standard scaler.

Step 6: data balancing

- To overcome the imbalance dataset use oversampling method like SMOTE.

Step 7: Data Splitting

- Training (80%)
- Testing (20%)

Step 8: classification model

- Apply gradient boost and LGBM method

Step 9: Model Training

- Train the model of the preprocess dataset.

Step 10: Model Evaluation

- Evaluation metrix like f1-score, recall, accuracy, and precision.

Step 11: predict outcome

Finish!!!!!

4. Results & Discussions

In this section, describe the dataset decription, and analysis of the dataset. The result of dataset analysis are present in this part.

4.1. Dataset Description

A BCWD may be found in the UCI ML repository. The primary topic matter of this collection is medicine and health. There are 569 occurrences and 30 characteristics in this collection. A digital image acquired by FNA of a breast lump is utilised to compute features. They provide details about a characteristic of a cell nuclei seen in an image. A comprehensive search inside the space of 1-4 features & 1-3 separation planes yielded relevant features. the genuine linear programme that was utilised to create the three-dimensional separation plane. An UW CS file transfer protocol server provides access to this database as well. Ten real-valued characteristics for every cell nucleus in this dataset are calculated. Creative Commons Attribution 4.0 is the licence for this dataset.

4.2. EDA (Exploratory Data Analysis)

A technique to data summarization known as EDA involves identifying the data's primary properties and then using appropriate representations to display them. EDA provides a concise overview of the data collection, including its size, kinds, missing data, and columns and rows. Find and fix missing data, incorrect data types, and incorrect values; eliminate inaccurate data. Bar graphs, histograms, or box plots are the visual representations of data distribution provided by EDA. Find the relationships (correlations) between the variables and show them on a heat map. Dataset information with an in-depth analysis is presented in the following graphic representations:

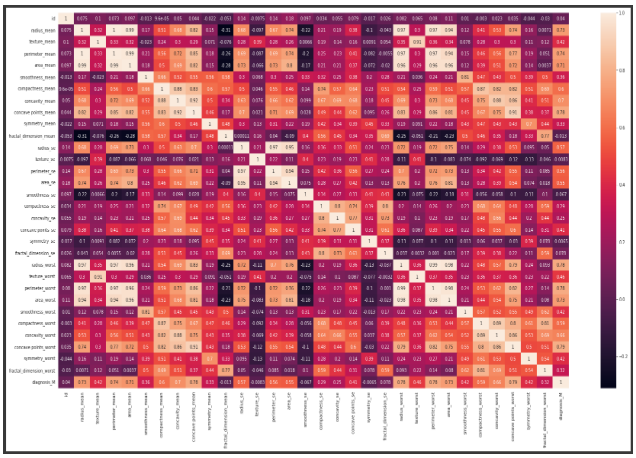

Fig. 3. Heatmap of Breast Cancer Wisconsin dataset

The above figure 3, represet a heatmap of BCWD. The x-axis displays data attributes, while the y-axis displays supplied data attributes in this figure.

4.3. Performance Measures

A forecasting effect of a model was assessed in this research using F1-score, recall, accuracy, and precision [34]. It is believed that all cases of malignant BC may be predicted due to the precision of medical diagnostics. The following performance measures are as follow:

4.3.1. Confusion Matrix

When true values are available for use in testing a classification model, a confusion matrix is a graphical representation of a prediction outcomes of a classifier. Below discussed both kind of classification i.e., binary and multi classification.

A binary classifier's confusion matrix is depicted in Fig. 5. True (1) and false (0) represent actual values, while negative (0) and positive (1) represent expected values. The expressions TP, TN, FP, and FN found in the confusion matrix are used to derive estimates of the possible classification models.

- **TP (True Positive):** In the confusion matrix, a data point is TP if and only if the expected outcome matches with the actual outcome.
- **FP (False Positive):** If a positive outcome is predicted but a negative consequence occurs in actual, this will be represented as a false positive in the confusion matrix.
- **FN (False Negative):** A false negative happens in the confusion matrix when a negative outcome is predicted but a positive outcome actually occurs.
- **TN (True Negative):** When an expected consequence is negative and an actual outcome is also negative,called TN.

Figure 4 depicts a confusion matrix used for the classification of four different classes. The classification of examples (instances) into four classes is known as four-class classification. Four classes include "Class A, Class B, Class C, and Class D".

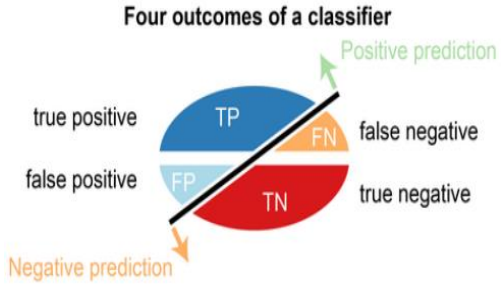

Fig. 4. Confusion Matrix of Binary Classification

4.3.2. Accuracy

Accuracy is a proportion of all correctly predicted samples in a whole sample, including both positive and negative samples. The equation is

$$Accuracy = \frac{TP + TN}{TP + Fp + TN + FN} \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

4.3.3. Precision

Precision, defined as the proportion of positive predicted samples that are really positive samples, is computed using the formula.

$$Precision = \frac{TP}{TP + FP} \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

4.3.4. Recall

The true positive rate is another name for recall. Recall for the first samples is the percentage of properly predicted positive samples in the total number of samples, and the equation is

$$Recall = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

4.3.5. F1 Score

The F1-score is calculated using the weighted harmonic average of recall and precision. An elevated F1-score indicates more effective classification outcomes. It is a thorough assessment indicator of external approaches. For the index F1-score, the formula is.

$$F1 - Score = \frac{2(Precision * Recall)}{\text{Precision} + \text{Recall}} \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

4.4. Experimental Analysis of proposed model with test dataset

In this part, provide the results of the suggested model using the test dataset. A result is as follow:

```
Testing results of Gradient Boost Model:

Testing Accuracy: 97.902
Testing F1 Score: 97.906
Testing Recall Score: 97.902
Testing Precision: 97.997
```

Fig. 5. Testing result of GB model

The above figure 5 represent the testing results of gradient boost model. The model testing performance is (accuracy is 97.902, precision is 97.997, recall is 97.902, and f1-score is 97.906.

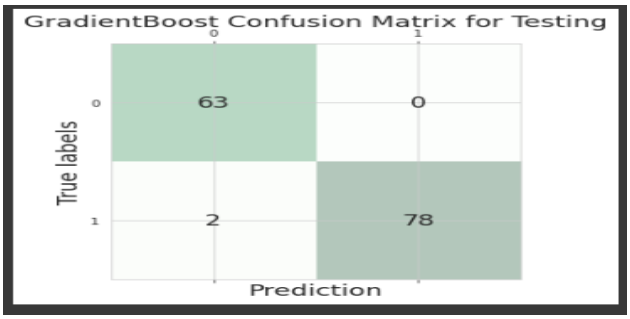

Fig. 6. Confusion metrix of GB model

In this figure 6 present the confusion metrix of gradient boost model. Class 1 includes both the TP and FP. Class 0 signifies both TN and FN. In this case, TP is 78, FP is 2, FN is 0, and TN is 63.

```
Classification report of Gradient Boost model for testing:

              precision    recall  f1-score   support

           0       0.97      1.00      0.98        63
           1       1.00      0.97      0.99        80

    accuracy                           0.99       143
   macro avg       0.98      0.99      0.99       143
weighted avg       0.99      0.99      0.99       143
```

Fig. 7. Classification report of GB model

The above figure 7 shows the classification report of gradient boost model. The characteristics of Class 0 are as follows: support value of 63, recall of 1.00, precision of 0.97. Class 1 consists of the following measures: support value of 80, recall of 0.97, and f1-score of 0.99. A weighted avg support value is 143.

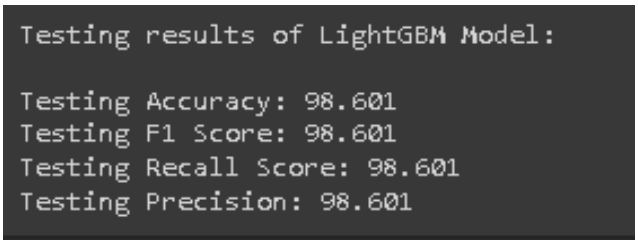

Fig. 8. Test result of LightGBM model

The above figure 8 represent the testing results of LightGBM model. The model testing performance is (accuracy is 98.601, precision is 98.601, recall is 98.601, and f1-score is 98.601.

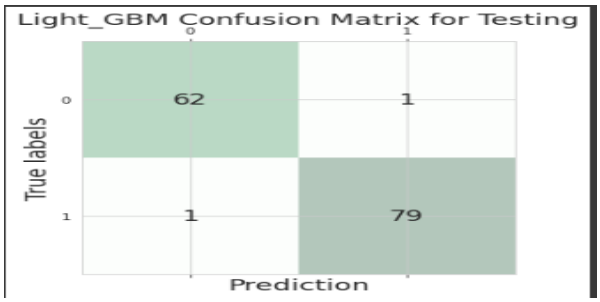

Fig. 9. Confusion metrix of lightGBM model

In this figure 9 present the confusion metrix of lightGBM model. Class 1 represent the TP and FP. Class 0 represent the FN and TN. The value of TP is 79, FP is 1, FN is 1, TN is 62

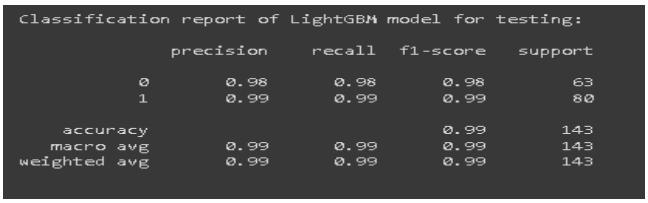

Fig. 10. Classification report of LightGBM model

Figure 10 above displays the LightGBM model's classification report. Class 0 should be given the following values: 63 for support, 0.98 for precision, and 0.98 for recall. Here are the Class 1 metrics: With a precision of 0.99 and a recall of 0.99, the support value is 80. The average weighted support value is 143.

4.5. Comparison Between Base and Proposed Models

Provide an explanation of the basic composition in this part and propose models using tables and graphs.

Table 2. Comparision table of proposed and base model

Parameters in (%)	Base Model		Proposed Models	
	XGBoost	Random forest	Gradient boost	LGBM
Accuracy	95.90	97.07	97.90	98.60
Precision	95.90	95.90	97.99	98.60
Recall	95.90	95.90	97.90	98.60
F1-score	95.89	95.86	97.90	98.60

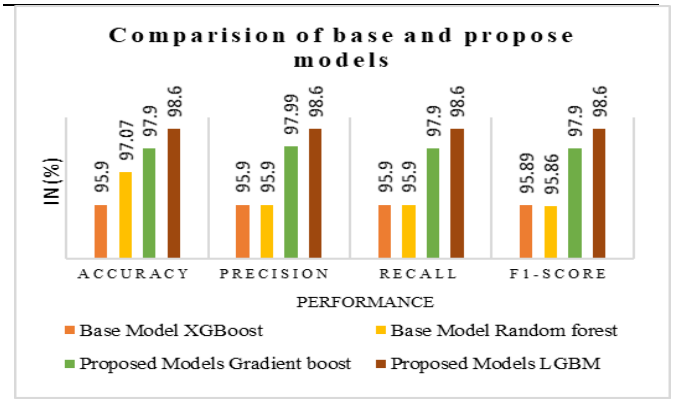

Fig. 11. Comparision graph of base and propose model

Figure 11 above displays a model comparison among a base and propose models. All of a proposed model outperform the Base Model on important performance metrics such as F1-score, Accuracy, Precision, and Recall. LGBM achieve higher performance is 98.6%. Insights for model selection are provided by this comparison study, which highlights the significant advantages of using complex algorithms to optimise prediction results.

5. Conclusion and Future Work

Experimental studies conducted recently have shown the effectiveness of machine learning approaches in building high-performance automated systems for cancer early diagnosis, therapy, and detection. In this work, they looked at multiple ML methods for identifying breast cancer. They conducted a comparison study between Random Forest, XGBoost, LightGBM, and gradient boost. It was observed that achieved achieved a higher efficiency 98.6%, whereas gradient boost achieved 97.9, XGBoost achieved 95.9, and random forest achieve 97.07%. Consequently, the early diagnosis and prognosis of different forms of cancer will depend heavily on supervised ML algorithms. While the current study underscores the considerable importance of machine learning methodologies in cancer detection, there remains room for additional development and expansion of the current research. In the foreseeable future, there will undoubtedly be proposals for more efficient and effective models, as new concepts, proposals, and suggestions continue to emerge in this field. Additional frameworks based on deep learning are also under our development.

References

[1] J. E. T. Akinsola, M. A. Adeagbo, and A. A. Awoseyi, "Breast cancer predictive analytics using supervised machine learning techniques," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2019, doi: 10.30534/ijatcse/2019/70862019.

[2] A. Brédart *et al.*, "Clinicians' use of breast cancer risk assessment tools according to their perceived importance of breast cancer risk factors: an international survey," *J. Community Genet.*, 2019, doi: 10.1007/s12687-018-0362-8.

[3] P. Maas *et al.*, "Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States," *JAMA Oncol.*, 2016, doi: 10.1001/jamaoncol.2016.1025.

[4] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, 2019, doi: 10.1148/radiol.2019182716.

[5] W. L. Bi *et al.*, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA. Cancer J. Clin.*, 2019, doi: 10.3322/caac.21552.

[6] T. Yanes, M. A. Young, B. Meiser, and P. A. James, "Clinical applications of polygenic breast cancer risk: A critical review and perspectives of an emerging field," *Breast Cancer Research*. 2020. doi: 10.1186/s13058-020-01260-3.

[7] H. Behravan, J. M. Hartikainen, M. Tengström, V. – M Kosma, and A. Mannermaa, "Predicting breast cancer risk using interacting genetic and demographic factors and machine learning," *Sci. Rep.*, 2020, doi: 10.1038/s41598-020-66907-9.

[8] K. Y. Obiwusi, Y. O. Olatunde, G. K. Afolabi, A. Oke, A. M. Oyelakin, and A. Salami, "Evaluating the Performance of Supervised Machine Learning Algorithms in Breast Cancer Datasets," *ASEAN J. Sci. Eng.*, 2023, doi: 10.17509/ajse.v3i2.46152.

[9] M. Murugesan, M. Santhosh, S. K. T, M. Sasiwarman, and I. Valanarasu, "International Journal of Advanced Trends in Computer Science and Engineering Securing ATM Transactions using Face Recognition," vol. 9, no. 2, pp. 1295–1299, 2020.

[10] N. Das, J. Borah, and K. Sarmah, "Diagnosis and Classification of Breast Cancer Using Multiple Machine Learning Algorithms," 2023. doi: 10.1109/InCACCT57535.2023.10141796.

[11] R. H. Khan, J. Miah, M. M. Rahman, and M. Tayaba, "A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer," 2023. doi: 10.1109/CCWC57344.2023.10099106.

[12] Jamal, J. H. Antor, R. Kumar, and P. Rani, "Breast Cancer Prediction Using Machine Learning Classifiers," 2022. doi: 10.1109/ICAST55766.2022.10039656.

[13] C. Roy, I. Mazumder, S. Debdas, S. Samanta, and S. S. Roy, "Framework for Breast Cancer Diagnosis Using Machine Learning and IoT," 2022. doi: 10.1109/ICEEICT53079.2022.9768469.

[14] S. Anklesaria, U. Maheshwari, R. Lele, and P. Verma, "Breast Cancer Prediction using Optimized Machine Learning Classifiers and Data Balancing Techniques," 2022. doi: 10.1109/ICCUBEA54992.2022.10010783.

[15] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "Breast Cancer Prediction Using Long Short-Term Memory Algorithm," 2022. doi: 10.1109/CINE56307.2022.10037258.

[16] H. Sharma, P. Singh, and A. Bhardwaj, "Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques," 2022. doi: 10.1109/ESCI53509.2022.9758188.

[17] H. Sami, M. Sagheer, K. Riaz, M. Q. Mehmood, and M. Zubair, "Machine Learning-Based Approaches for Breast Cancer Detection in Microwave Imaging," 2021. doi: 10.23919/USNC-URSI51813.2021.9703518.

[18] V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," 2020. doi: 10.1109/C2I451079.2020.9368911.

[19] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018. doi: 10.1109/ETECHNXT.2018.8385355.

[20] M. S. Harinishree, C. R. Aditya, and D. N. Sachin, "Detection of Breast Cancer using Machine Learning Algorithms - A Survey," 2021. doi: 10.1109/ICCMC51019.2021.9418488.

[21] M. E. Gamil, M. Mohamed Fouad, M. A. Abd El Ghany, and K. Hoffinan, "Fully automated CADx for early breast cancer detection using image processing and machine learning," 2018. doi: 10.1109/ICM.2018.8704097.

[22] Vikas Chaurasia and Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability by Vikas Chaurasia, Saurabh Pal :: SSRN," *Int. J. Comput. Sci. Mob. Comput. IJCSMC*, 2014.

[23] A. U. Haq *et al.*, "Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised

Feature Selection Techniques," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3055806.

[24] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018. doi: 10.1109/CTEMS.2018.8769187.

[25] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," *SN Comput. Sci.*, 2020, doi: 10.1007/s42979-020-00305-w.

[26] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowl. Eng. Rev.*, 2019, doi: 10.1017/S026988891800036X.

[27] J. Li *et al.*, "Feature selection: A data perspective," *ACM Computing Surveys*. 2017. doi: 10.1145/3136625.

[28] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, 2009, doi: 10.1142/S0218001409007326.

[29] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," 2005. doi: 10.1007/11538059_91.

[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, 2001, doi: 10.1214/aos/1013203451.

[31] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, 2021, doi: 10.1007/s10462-020-09896-5.

[32] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms," *Mathematics*, 2020, doi: 10.3390/MATH8050765.

[33] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud Detection in Banking Data by Machine Learning Techniques," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2022.3232287.

[34] S. Libesman *et al.*, "An individual participant data meta-analysis of breast cancer detection and recall rates for digital breast tomosynthesis versus digital mammography population screening," *Clinical Breast Cancer*. 2022. doi: 10.1016/j.clbc.2022.02.005.