Inteligencia Artificial 28(76), 149-165
doi: 10.4114/intartif.vol28iss76pp149-165

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Breast Cancer Prediction using Stacking Models & Hyperparameter Tuning

Rahul Karmakar[1], Akhil Kumar Das[2*], Debapriya Sarkar[3], Saroj Kumar Biswas[4], Ardhendu Mandal[5], Arijit Bhattacharya[6]

[1] Department of Computer Science, Assistant Professor, The University of Burdwan, Burdwan, West Bengal, India.
[2*] Department of Computer Science, Assistant Professor, Gour Mahavidyalaya, Malda-732142, West Bengal, India.
[3] Department of Computer Science, The University of Burdwan, Burdwan, West Bengal, India.
[4] Department of Computer Science and Engineering, Associate Professor, National Institute of Technology, Silchar, Assam, India.
[5] Department of Computer Science and Technology, Associate Professor, University of North Bengal, Darjeeling, India.
[6] Department of Computer Science, Assistant Professor, Gour Mahavidyalaya, Malda-732142, West Bengal, India.

[1] rkarmakar@cs.buruniv.ac.in
[2*] dasakhi@gmail.com
[3] debapriyasarkar.bardhaman@gmail.com
[4] bissarojkum@yahoo.com
[5] am.csa.nbu@gmail.com
[6] barijit@hotmail.com

Abstract This paper explores the application of stacking models for breast cancer detection, integrating key techniques such as data balancing, hyperparameter tuning, and feature selection. We implemented five different stacking configurations. Initially, Logistic Regression (LR) was used as the meta-classifier, while the base estimators included Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) classifiers. In the second configuration, we reversed the roles: DT acted as the meta-classifier, with SVM, KNN, RF, and LR serving as the base estimators. In a third setup, SVM was used as the meta-classifier, with DT, LR, KNN, and RF as the base learners. Fourth, we implemented KNN as the stacking classifier, with LR, DT, SVM, and RF as the base estimators. Finally, in the fifth configuration, RF was the meta-classifier, supported by LR, DT, KNN, and SVM as base learners. The evaluation of stacking models was conducted in five phases, starting with a baseline with no adjustments, followed by applying data balancing alone, then adding hyperparameter tuning, applying Chi-square feature selection with data balancing, and finally using correlation-based feature selection with data balancing, all systematically excluding certain elements to analyze their individual impact. Among all cases, the stacking model with LR delivers the best performance, achieving an accuracy of 97.63%, precision of 97.68%, recall of 97.63%, and an F-measure of 97.63%, showcasing its exceptional reliability and balanced effectiveness. All models were evaluated using 10-fold cross-validation.

Keyword: Breast Cancer, Feature Selection, Hyperparameter Tuning, Cross Validation

1 Introduction

Breast Cancer (BC) is the primary cause of death. BC is the most common cancer in women globally. It is frequently characterized by the unchecked proliferation of breast cells, which can result in lumps or tumors that can be seen on X-rays or other medical imaging tests. A major difficulty in diagnosing BC is differentiating between benign and malignant tumors [1],[2]. In medical data analysis, predicting BC is a difficult task. For the purpose of guiding future treatments, early diagnosis of this malignancy is crucial. For pathologists and doctors to make decisions and distinguish between benign and malignant tumors, they needed certain automated technologies. A number of techniques can be used to diagnose this illness, including thermography, ultrasonography, mammography, breast biopsy, and fine-needle aspiration cytology. Mammography has been the

ISSN: 1137-3601 (print), 1988-3064 (on-line)

most common method up to this point for identifying BC; but, in certain situations, it is not enough, and a biopsy is necessary before a diagnosis can be made [3],[4].

This process is not only time-consuming but also susceptible to human error. This may prevent patients from receiving the necessary treatments in a timely manner. Therefore, it is essential to develop a Machine Learning (ML) based system that can detect BC at an early stage using clinical symptoms. By distinguishing BC from other common illnesses, such as a usual fever, this method makes it possible to diagnose and treat the condition in a timely manner. Because of this, the medical expert system not only cuts down on the expenses and the amount of time that are connected with pathological diagnosis, but it also greatly minimizes the danger of individuals passing away [5],[6]. Furthermore, ML greatly facilitates decision-making and diagnosis based on data gathered by the medical industry [7], [8]. Numerous studies demonstrate the value of ML techniques in decision-making related to BC prediction [9]. Every expert system has pros and cons of its own. Many systems struggle with class imbalance, outliers, data pre-processing, and feature selection. To overcome the problem, this study proposes an ML-based system for early breast cancer detection using symptoms, aiming to improve accuracy, reduce treatment costs, and enable early intervention.

This study's objectives focus on the following aspects:

- This paper aims to evaluate the effectiveness of stacking models for breast cancer detection by integrating advanced techniques such as data balancing, hyperparameter tuning, and feature selection. The study systematically compares five stacking configurations with different meta-classifiers and base learners, analyzing their performance across various evaluation phases to identify the optimal configuration for reliable and accurate breast cancer detection.
- This study uses ML methods to identify important clinical variables for diagnosing malignant and benign breast tumors. It also evaluates how these variables affect prediction accuracy in BC patients. The performance of the methods will be tested using 10-fold CV, measuring accuracy, recall, precision, and F-measure.

The key contributions of this study are centered on the following aspects:

- The KMeansSMOTE method is used to balance the BC dataset. This approach creates a well-balanced dataset, improving the model's ability to handle different data distributions. Comparing the classification and prediction accuracy of different stacking methods for BC.
- Address the issue of irrelevant features by using two feature selection methods, Chi-square and correlation techniques, to extract important features that improve data representation.
- Employ hyperparameter tuning to identify the optimal values for the parameters of the model.
- The efficacy of the proposed strategy in relation to existing state-of-the-art techniques.

This paper is structured as follows. Six sections comprise the study: In section 2, a review of the literature is provided, section 3 displays the recommended methodology as well as pre-processing, in section. 4, the outcome and analysis are explained and section 5 concludes with suggestions for additional study.

2 Literature Review

The following review work was carried out utilizing WBCD dataset. In a research paper "Wisconsin breast cancer diagnostic dataset" was used for analysing. The Extreme Gradient Boosting (XGBoost) machine learning algorithm, which is well-known for its efficacy and efficiency in classification tasks, was used in the study. With an accuracy of 94.74% and a recall of 95.24%, the research showed remarkable results [10]. When evaluating the effectiveness of several classifiers, such as KNN, RF, DT, LR, and SVM, a research was carried out with the WBCD dataset as the basis for the investigation. It was determined that RF had the highest accuracy, which was

96.44% [11]. In another study, they presented four ML algorithms including XGBoost, RF, LR, and K-NN for BC prediction. XGBoost performed better when compared to the other methods, according to the study. It achieved F1-scores of 0.980, whereas accuracy, precision, and recall score were recorded at 1.00, 0.960, and 0.974, respectively, utilizing 80% training data and 20% test data [12]. In this research paper [13], stand-alone ML classifiers such as NB, LR, SVM, K-NN, and DT, along with ensemble ML classifiers like RF, AdaBoost, and XGBoost, were utilized for breast cancer detection. To evaluate model performance, metrics including accuracy, precision, AUC, and recall were employed. Among all models, DT and XGBoost achieved the highest accuracy of 97%, with XGBoost also attaining the highest AUC of 0.999. In this study [14], an LR classifier was employed for BC detection. The evaluation metrics included accuracy, precision, recall, and F1 score. Using LR with 10-fold CV, the model achieved an accuracy of 96.6%. In this paper [15], a ML approach utilizing the LR classifier was proposed for breast cancer detection. The model achieved an accuracy of 96% using 10-fold cross-validation. This paper [16] introduced a novel approach by proposing a model named “SELF” for BC detection using two datasets: the BreakHis dataset with 82 instances and the WBCD dataset [59]. The study utilized five classifiers: Extra-Trees, RF, AdaBoost, GBoost, and K-NN. The proposed model achieved an accuracy of 95% on the BreakHis dataset and 99% on the WBCD dataset. The "Breast Cancer Wisconsin Diagnostic (WBCD)" dataset, comprising 569 patients (62.74% benign and 37.26% malignant) with 33 patient attributes was utilized in a study. Many classification methods including DT, ANN, rough set, NB, SVM, and KNN were applied with 10-fold CV. With 96.79% accuracy, the SVM (SMO) classifier has the highest accuracy. Based on the classification results of the six approaches, SVM (SMO) outperforms the other five algorithms for the chosen dataset [17]. This research paper presented on a supervised ML based system for BC classification employing FS, PCA, grid search for hyperparameter tuning, and CV. Using majority voting, two ensemble models and seven ML classifiers (ANN, KNN, SVM, DT, RF, XGBoost, and AdaBoost) were concatenated and stacked using logistic regression S-LR. Wisconsin and Mass mammography (MM) were used. The results showed that the XGBoost model for the MM dataset had the highest recall, at over 96 percent. The AdaBoost and S-LR models outperformed the others with a recall of 95.35 percent for the WBCD [18]. Using the WBCD dataset, a study evaluated the performance of many boosting classifiers. The study's findings indicate that AdaBoost and XGBoost performed more accurately than Gradient Boosting. AdaBoost got the highest accuracy of 98.6% after implementing hyperparameter adjustment, followed by Gradient Boosting at 95.8% and XGBoost at 97.2 percent. The results showed that XGBoost and AdaBoost had the highest accuracy rates, at 98.60 percent and 97.20 percent, respectively. GB also performed well, receiving an accuracy rating of 95.80 percent. When compared to previous research findings on the WBCD dataset, XGBoost and AdaBoost fared better than the other classifiers in this dataset. When compared to other previous studies, which showed accuracy rates of 95.34 and 97.34 percent, GB also performed well [19]. SVM, RF, LR, DT, and KNN are the five main algorithms that were employed in this work on the WBCD dataset to evaluate different outcomes based on the AUC, accuracy, sensitivity, precision, and confusion matrix. An exact comparison of our models shows that SVM outperforms all other algorithms, achieving superior accuracy of 97.2 percent, precision of 97.5 percent, and AUC of 96.6 percent. In conclusion, SVM achieved the highest degree of precision and accuracy and demonstrated efficacy in the prediction and diagnosis of breast cancer [20]. Data from the UCI ML Repository was used in a study. In this case, roughly 67% of the data were utilized to train the model and 33% were used to test it. Two Ensemble ML algorithms, RF & XGBoost, were put into practice. C4.5 or J48 is the classifier that RF use. We have obtained accuracy by using the algorithms, such as 74.73% accuracy for Random Forest and 73.63% accuracy for XGBoost [21]. A study made use of the WBC and WBCD databases. BC was classified and diagnosed using six ML algorithms: ANN, KNN, SVM, DTs, LR, and Bayes Network (BN). Wrapper approaches were used for FS, combining the classifier subset evaluator methodology with the best first search method to improve the classification accuracy. The feature selection strategy improved WBC (from 97.2818 to 97.4249) and WBCD (from 95.2548 to 96.1336) classification accuracy for some classifiers (like Bayes net); however, it decreased WBC (from 97.2818 to 95.279) & WBCD (97.891 to 97.3638) classification accuracy for other classifiers (like SVM). The outcome demonstrated that the best model for WBC breast cancer classification was the Bayes Network with FS, while the most potent model for WBCD was the SVM without FS [22].

Table 1: Summary of literature survey

Study	Dataset	Methods	Accuracy (Best)	Key Finding	Limitation
Rahmanul Hoque et al. (2024)[10]	WBCD[23]	XGBoost	94.74%	Several FS techniques, ML classifiers & hyperparameter tuning were used.	No FS and no scaling
Rahul Karmakar et al.(2023)[11]	WBCD	KNN, RF, DT, LR, and SVM	96.44%	Various Train-Test splits were used.	No balanced dataset and no scaling
Hua Chen et al. (2023)[12]	WBCD	XGBoost, RF, LR, and K-NN	97.4%	PCC was employed	CV is not used. Less comparisons of previous work.
Varsha Nemade et al. (2023) [13]	WBCD	NB, LR, SVM, K-NN DT, RF, AdaBoost, and XGBoost	97%	No FS is employed and many classifiers were employed.	No proper pre-processing. Unbalanced dataset. No CV is used.
Annisa Maulidia et al. (2023) [14]	WBCD	LR	96.5%	No FS is employed. Only one classifier was employed	10-fold CV is used. No scaling. Unbalanced dataset. No comparisons of previous work.
Saheb Karan et al. (2023) [15]	WBCD	LR	95.75%	10-fold CV is used. But No FS	No scaling and Over fitting problem.
Amit Kumar Jakhar et al. (2023) [16]	WBCD	EXTR, RF, AdaBoost, GBoost, and K-NN	95%	Five classifiers were employed to obtain the results	SMOTE utilized, FS Utilized, Outlier detection and removal not utilized. 10-fold CV utilized.
Ruchika Patel(2023)[17]	WBCD	SVM (SMO)	97.89%	Several methods and strategies were employed to gain a better outcome.	No FS and no data balancing, No comparisons of previous work.
Sara Laghmati et al. (2023)[18]	WBCD & MM	S-LR	WBCD: 97.37% , and MM: 93.7%	Several FS strategies were employed.	No data balancing and Outlier is not considered.
Md. Mijanur Rahman et al. (2023) [19]	WBCD	AdaBoost	98.60%	Three ML classifiers were used.	No proper pre-processing, and no data balancing.
Mohammed Amine Naji et al. (2021)[20]	WBC [24]	SVM	97.2%	Several ML classifiers and stacking models were used.	No FS and no data balancing.
Sajib Kabiraj et al. (2020)[21]	UCI BCD	RF & XGBoost	74.73%	Many ML classifiers & hyperparameter tuning were used.	No data balancing and outlier is not considered.
Hajar Saoud et al. (2019)[22]	WBC & WBCD	Bayes Net	WBCD: 97.36%, and WBC : 97.42%	Six classifiers used to identify the BC and Wrapper FS techniques were used.	No data balancing and No comparisons of previous work.

Table 1 includes a summary of previous studies, their performance, databases used, machine learning techniques applied, publication year, and references. It also highlights the limitations and challenges of prior research.

The limitations are as follows:

- The issue of class imbalance has not been sufficiently addressed.
- Outlier detection remains inadequately tackled.
- Cross-validation, essential for validating system performance, was not implemented in all systems.
- The feature selection techniques lacked a systematic approach.

This proposed article addresses all the aforementioned concerns, except for outlier handling.

3 Proposed method and pre-processing

3.1 Proposed Method

In ML, stacking models are used to increase overall performance by combining the predictions of several basic models. This strategy, also known as stacked generalization, uses the strengths of many algorithms to lessen the biases, variation, and mistakes that individual models may have. In this study, we employed five distinct stacking model forecasting techniques to estimate the BC. LR was used as the stacking classifier in the beginning, and the estimators were DT, SVM, KNN, and RF classifiers. Then, we chose SVM, KNN, RF, and LR as estimators and DT as the stacking model. We used DT, LR, KNN, and RF as estimators and SVM as a stacking classifier, LR, DT, KNN, and RF as estimators and KNN as a stacking model, following that, LR, DT, and KNN, SVM as estimators and RF as a stacking classifier. Without hyperparameter adjustment, FS, or data balancing, we evaluated our stacking models.

We evaluated our stacking model under various conditions to analyze its performance. Initially, the assessment was conducted without applying data balancing, hyperparameter tuning, or FS.

Data balance makes sure that each class in a dataset is reflected evenly, fixing problems where one class is over represented. It is very important for making models work better, especially when minority classes are important, like in medical diagnosis or scam discovery. Next, the model in this study was evaluated using data balancing alone, without applying FS or hyperparameter tuning. Hyperparameter tuning is the process of optimizing model parameters that are not learned during training to achieve better performance. Proper tuning significantly impacts a model's accuracy and efficiency. Following this, data balancing and hyperparameter tuning were applied, excluding FS. Subsequently, we incorporated data balancing and FS using the Chi^2 method, omitting hyper-parameter tuning. Finally, the model was tested with data balancing and FS using correlation, without hyper-parameter tuning.

The model was trained using 10-fold CV, improving reliability by evaluating it on different data subsets. This reduces overfitting, enhances generalization, and provides a more accurate estimate of its performance on unseen data. Using Figure 1, the workflow of the suggested method is demonstrated.

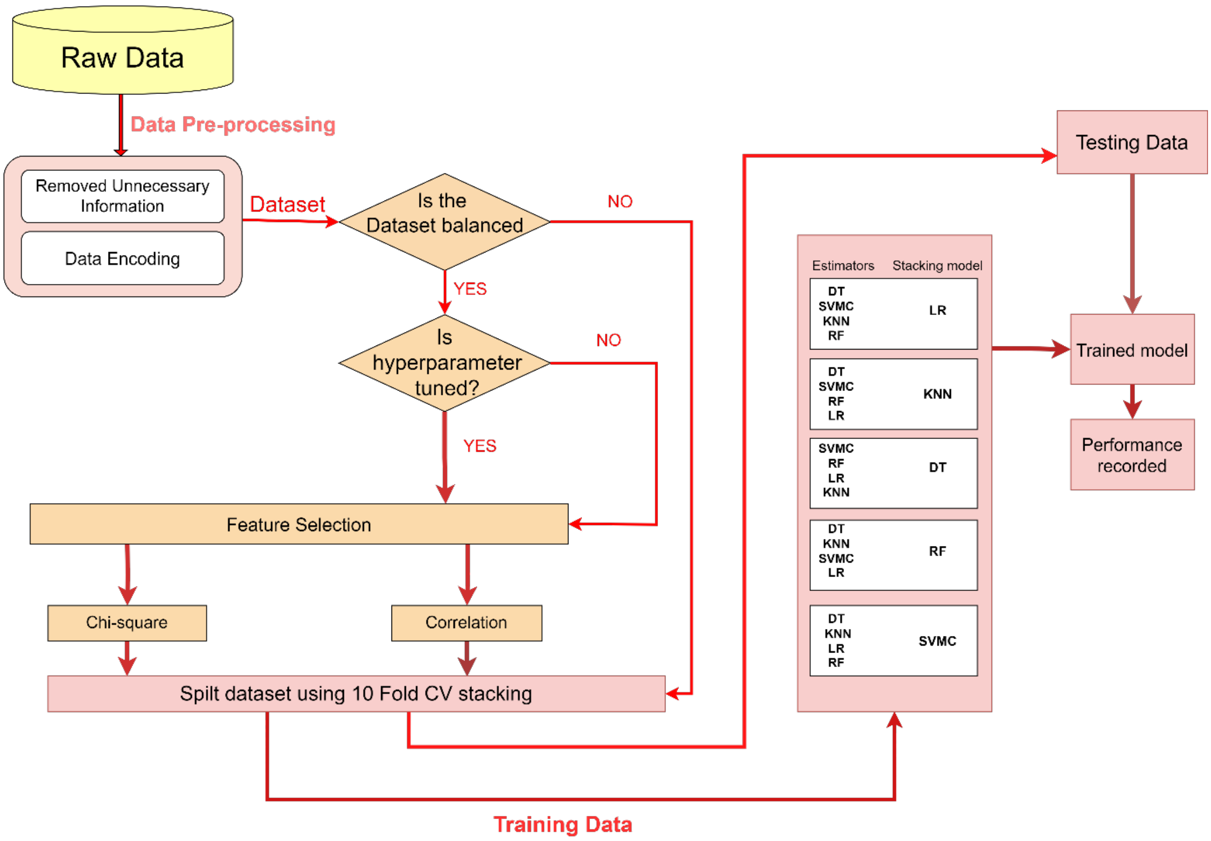

Figure 1. Describe the workflow of the work.

3.2 Dataset Collection and Description

We used the "Breast Cancer WISCONSIN (Diagnostic) [23]" dataset from Kaggle for our analysis. There are 33 characteristics and 569 instances in this dataset. There are '212' cases of malignancy and '357' benign cases in the sample. The designations 'Benign' and 'Malignant' are given to these two classes, respectively. The dataset contains no null or missing values. Table 2 shows visual distribution of data.

Table 2: Data distribution in a visual format

SL NO	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	
0	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	.
1	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	.
2	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	.
3	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	.
4	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	.
...	...	...	...	...	...	...	...	...	...	.
564	926424	M	21.56	22.39	142	1479	0.111	0.1159	0.2439	

										.
565	926682	M	20.13	28.25	131.2	1261	0.0978	0.1034	0.144	.
566	926954	M	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	.
567	927241	M	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	.
568	92751	B	7.76	24.54	47.92	181	0.05263	0.04362	0	.
569 rows × 33 columns										

3.3 Data pre-processing

The BC dataset is inaccurate, missing some patterns, incomplete, and prone to many errors. During the data pre-processing step, raw BC data is transformed into a format that is appropriate. The following are the steps involved in data preparation:

- **Removing the Irrelevant Features**
- **Level Encoding**
- **Data Balancing Using K-Means SMOTE Oversampling**
- **Feature Selection**

3.3.1 Removing the Irrelevant Features

The BC dataset contains two features that are irrelevant for diagnosing BC: 'Patient ID' and 'Unnamed: 32'. As part of the initial data preprocessing step, these two features are removed to ensure the dataset focuses only on relevant attributes for diagnosis. By removing irrelevant features, the proposed work becomes more efficient, interpretable, and reliable, ultimately leading to better outcomes.

3.3.2 Level Encoding

After removing the irrelevant features, the BC dataset contains 31 attributes of the object data type. Among these, the 'diagnosis' attribute, which is categorical, is transformed into numerical values: zero for benign tumors and one for malignant tumors. The data distribution following encoding is shown in table 3 with red color.

Table 3: Data distribution after encoding

SL NO	diagnosis	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	.
.	1	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	.
.	1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	.
.	1	19.69	21.25	130	1203	0.1096	0.1599	0.1974	.
.	1	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	.
.	1	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	.
.	...	...	...	...	...	...	...	...	.
.	1	21.56	22.39	142	1479	0.111	0.1159	0.2439	

									.
.	1	20.13	28.25	131.2	1261	0.0978	0.1034	0.144	.
.	1	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	.
.	1	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	.
.	0	7.76	24.54	47.92	181	0.05263	0.04362	0	.
569 rows × 31 columns									

3.3.3 Data Balancing Using K-Means SMOTE Oversampling

K-Means SMOTE (Synthetic Minority Oversampling Technique) is a data balancing technique that addresses class imbalance by generating synthetic samples for the minority class. It enhances traditional SMOTE by first clustering the minority class data using the K-Means algorithm. This clustering ensures that the synthetic samples are created within relevant sub-clusters, preserving the data's structure and improving the quality of oversampling. This approach is particularly effective in handling complex, imbalanced datasets by reducing the risk of overfitting and ensuring a more representative balance between classes. This study addressed an imbalance in the dataset, which contained ‘212’ malignant cases compared to ‘357’ benign cases. To mitigate this imbalance, the K-Means SMOTE technique was utilized for data balancing. After balancing the dataset, it contained ‘361’ malignant cases and ‘357’ benign cases.

3.3.4 Feature Selection (FS)

FS is the process of finding and choosing the most important features from a dataset in order to improve the performance of a ML model. It helps to minimize dimensionality, remove unnecessary or duplicated data, and improve computing performance. It avoids overfitting by focusing on the most important characteristics, simplifies the model, and enhances interpretability. In the BC dataset, which comprises 31 features, not all features are equally important for the analysis. To identify the most essential features, we applied two feature selection methods: the Chi-squared (Chi^2) technique and correlation analysis.

Chi^2 technique

This method assesses the relationship between each feature and the target variable. Each feature's significance can be determined by calculating the Chi^2 statistic and the corresponding p-value, after which the most significant features are chosen. Here, twenty-five features have been selected using the Chi^2 method. Table 4 displays the features selected using the Chi-squared (Chi^2) method.

Benefits:

- Capable of assessing how factors relate to each other.
- Ascertains the degree of difference between the observed and expected values.

Table 5: Features selected by correlation matrix

Total feature in dataset	After correlation FS: No of FS

'radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean','smoothness_mean', 'compactness_mean', 'concavity_mean','concave points_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst', 'diagnosis'.	'texture_mean', 'smoothness_mean', 'compactness_mean','concave points_mean', 'symmetry_mean', 'fractal_dimension_mean','texture_se', 'area_se', 'smoothness_se', 'compactness_se','concavity_se', 'concave points_se', 'symmetry_se','fractal_dimension_se', 'texture_worst', 'area_worst','smoothness_worst', 'compactness_worst', 'concavity_worst','concave points_worst', 'symmetry_worst', 'fractal_dimension_worst', 'diagnosis'

3.4 Methodology

The suggested approach used a stacking model to forecast the BC in five different ways. Firstly, we took LR as stacking classifier & DT, SVM, KNN, RF classifiers as estimators. Then we selected DT as stacking model & SVM, KNN, RF, LR as estimators. We took SVM as stacking classifier also & DT, LR, KNN, RF as estimators, KNN as stacking model & LR, DT, KNN, RF as estimators, thereafter RF as stacking classifier & LR, DT, KNN as estimators.

We evaluated our stacking model under various conditions:

- **Without Data Balancing, Hyperparameter Tuning, or Feature Selection (FS):** The initial assessment was conducted without applying any of these techniques.

- **With Data Balancing but Without FS or Hyperparameter Tuning:** The model was then evaluated using data balancing alone.

- **With Data Balancing and Hyperparameter Tuning:** Next, the model was assessed with data balancing and hyperparameter tuning, excluding feature selection.

- **With Data Balancing and FS (Chi2) but Without Hyperparameter Tuning:** Subsequently, the evaluation included data balancing and FS using the Chi2 method while omitting hyperparameter tuning.

- **With Data Balancing and FS (Correlation) but Without Hyperparameter Tuning**: Finally, the model was tested with data balancing and FS using correlation, without hyperparameter tuning.

4 Result and Analysis

This section aims to evaluate the prediction capabilities of the suggested methodology, employing the method 10-fold CV. The trials were carried out using Python as the programming language. The WBCD dataset was used in a number of experiments, and the results were carefully studied to find the real improvements that would help make the suggested model for adjustment better.

4.1 Generation results were obtained using a 10-fold CV without FS, Hyperparameter tuning and Data balancing

The empirical results obtained for the WBCD dataset are shown in Table 6. The table compares the performance of stacking models with different base estimators and meta-learners in terms of Accuracy, Precision, Recall, and F-Measure (all in percentages). Each stacking model combines four base estimators, with the meta-learner varying across models. Using DT, SVM, KNN, and RF as base estimators and LR as the meta-learner yielded the best performance, achieving 95.96% accuracy, the highest among all configurations. Figure 3 illustrates the accuracy of various stacking models in the absence of data balancing and feature selection methods, providing a clearer understanding of their baseline performance.

Table 6: Describe the results that, without data balancing, FS, or hyperparameter tuning

Estimators	Stacking Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
DT, SVM, KNN, and RF	LR	95.96	96.06	95.96	95.93
DT, SVM, RF, and LR	KNN	95.61	95.75	95.61	95.55
SVM, RF, LR and KNN	DT	94.72	94.89	94.72	94.71
DT, KNN, SVM, and LR	RF	94.91	95.09	94.91	94.89
DT, KNN, LR, and RF	SVM	95.44	95.59	95.44	95.39

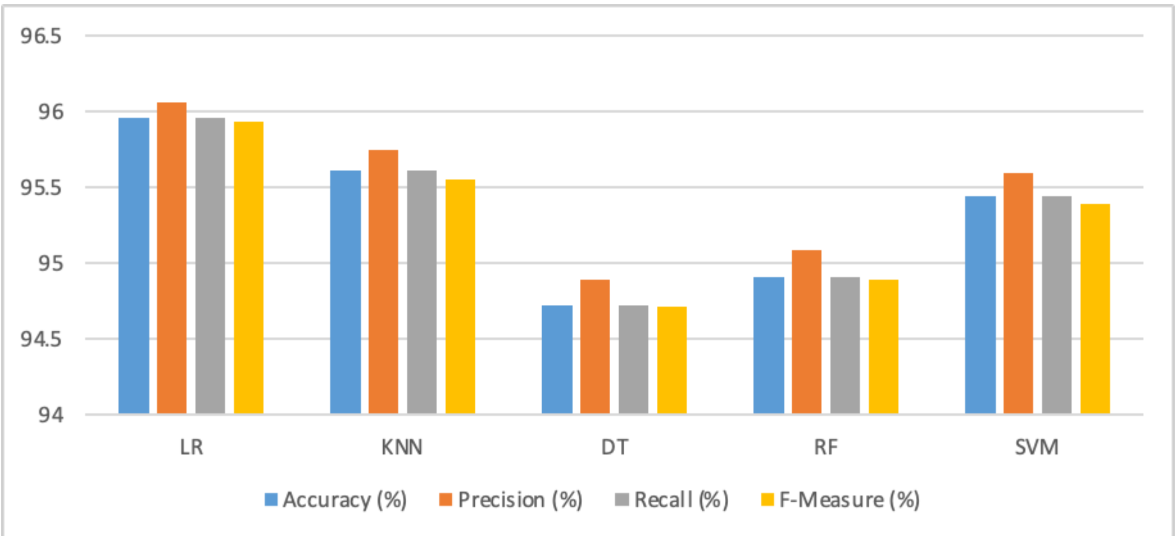

Figure 3. Visual distribution of accuracy across stacking models without data balancing, FS, or hyperparameter tuning.

4.2 Generation results were obtained using a 10-fold CV with Data balancing & without FS, Hyperparameter tuning

The empirical results for the WBCD dataset are presented in Table 7. The table summarizes the performance of various stacking models in a classification task. The models use combinations of base estimators (DT, SVM, KNN, RF, and LR) and a meta-classifier. Accuracy, precision, recall, and F-measure percentages indicate the effectiveness of each stacking model. The highest performance (96.94% accuracy, 97.10% precision, 96.94% recall, and 96.93% F-measure) is achieved when DT, SVM, KNN, and RF are stacked with LR as the meta-classifier. Comparable results (96.94% accuracy, 97.06% precision, and 96.94% for recall and F-measure) are obtained when DT, KNN, SVM, and LR are stacked with RF as the meta-classifier. The lowest performance (95.40% across all metrics) is observed when SVM, RF, LR, and KNN are stacked with DT as the meta-classifier.

This analysis highlights the impact of different combinations of base estimators and meta-classifiers on model performance. Data imbalance can have a significant impact on model performance, particularly in classification tasks, by causing the model to favor the majority class. Balancing the dataset ensures that the model treats all classes equally, preventing bias towards the majority class. In the case of the WBCD dataset, applying data balancing led to an overall improvement in model performance, allowing for more accurate and fair classification results across all classes. Figure 4 visually depicts the accuracy of several stacking models with data balance, excluding FS and hyperparameter adjustment.

Table 7: Describe the results that, data balancing, without FS, or hyperparameter tuning

Estimators	Stacking Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
DT, SVM, KNN, and RF	LR	96.94	97.10	96.94	96.93
DT, SVM, RF, and LR	KNN	96.53	96.69	96.53	96.52
SVM, RF, LR and KNN	DT	95.40	95.56	95.40	95.40
DT, KNN, SVM, and LR	RF	96.94	97.06	96.94	96.94
DT, KNN, LR, and RF	SVM	96.80	96.97	96.80	96.80

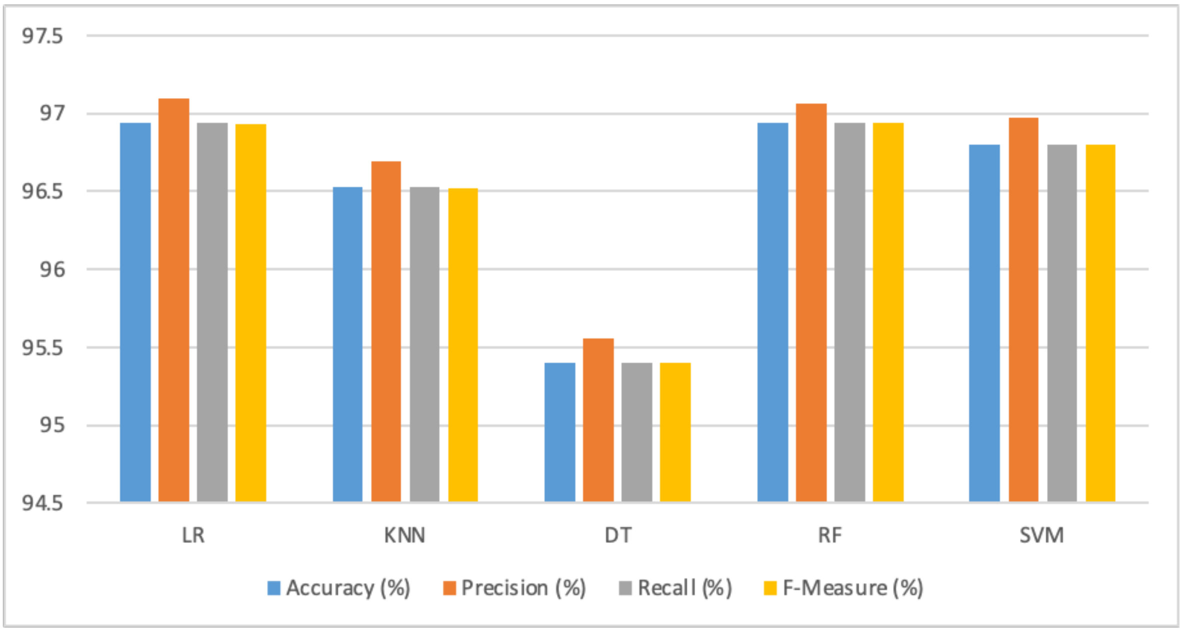

Figure 4. Visual distribution of different stacking models with data balancing & without FS & hyperparameter tuning.

4.3 Generation results were obtained using a 10-fold CV with Data balancing & hyperparameter tuning, without FS

Table 8 presents the empirical results obtained for the WBCD dataset. The table presents the performance metrics of different stacking models in a classification task, where various combinations of base estimators (DT, SVM, KNN, RF, and LR) are used with different meta-classifiers. The highest accuracy (97.63%) is achieved when DT, SVM, KNN, and RF are stacked with LR as the meta-classifier, alongside precision, recall, and F-measure values of 97.68%, 97.63%, and 97.63%, respectively. This analysis highlights that LR as meta-classifiers generally deliver superior results in stacking models. When using data balancing and hyperparameter tuning, make sure to use the right evaluation metrics to measure performance, especially for imbalanced datasets. Applying data balancing and hyperparameter tuning to the WBCD dataset resulted in a significant improvement in model performance, enhancing the accuracy and fairness of classification across all classes. Figure 5 illustrates the accuracy of different stacking models with data balancing, Chi^2 FS & without hyperparameter tuning.

Table 8: Describe the results that data balancing & hyperparameter tuning, without FS

Estimators	Stacking	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)

	Model				
DT, SVM, KNN, and RF	LR	97.63	97.68	97.63	97.63
DT, SVM, RF, and LR	KNN	94.85	95.23	94.85	94.84
SVM, RF, LR and KNN	DT	95.69	95.84	95.69	95.69
DT, KNN, SVM, and LR	RF	97.36	97.47	97.36	97.35
DT, KNN, LR, and RF	SVM	96.80	96.96	96.80	96.80

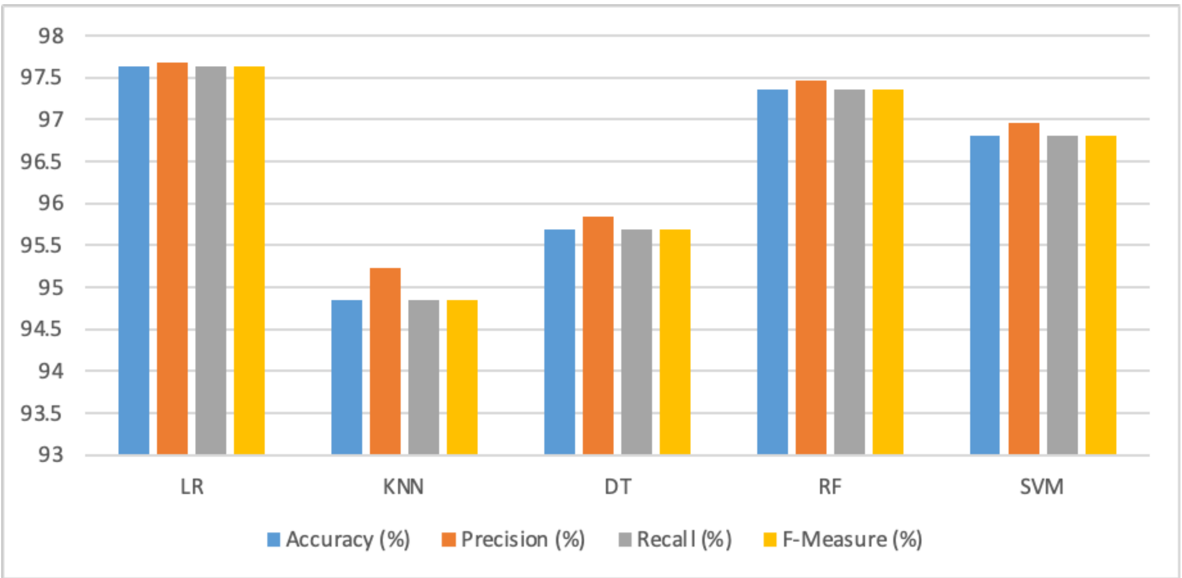

Figure 5. Visual distribution of different stacking models with data balancing & hyperparameter tuning, without FS

4.4 Generation results were obtained using a 10-fold CV with data balancing, FS (Chi2) & without hyperparameter tuning

The empirical results obtained for the WBCD dataset are shown in Table 9. Performance-wise, the stacking model LR and DT, SVM, KNN, RF classifiers as estimators performed better than other ways in the 10-fold CV. The table delineates the performance of several models inside a stacking framework, evaluated using four metrics: accuracy, precision, recall, and F-measure. The stacking model with estimators DT, KNN, SVM, RF, and LR as base estimator achieved the highest accuracy (96.94%), precision (97.09%), recall (96.94 %), and F-Measure (96.94 %) among the stacking framework. Figure 6 visually illustrates the accuracy of different stacking models with data balancing, Chi^2 FS & with hyperparameter tuning.

Table 9: Describe the results that Data balancing and FS (Chi^2) & without hyperparameter tuning

Estimators	Stacking Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
DT, SVM, KNN, and RF	LR	96.94	97.09	96.94	96.94
DT, SVM, RF, and LR	KNN	96.11	96.42	96.11	96.10
SVM, RF, LR and KNN	DT	95.41	95.54	95.41	95.41
DT, KNN, SVM, and LR	RF	96.52	96.67	96.52	96.52
DT, KNN, LR, and RF	SVM	96.38	96.59	96.38	96.38

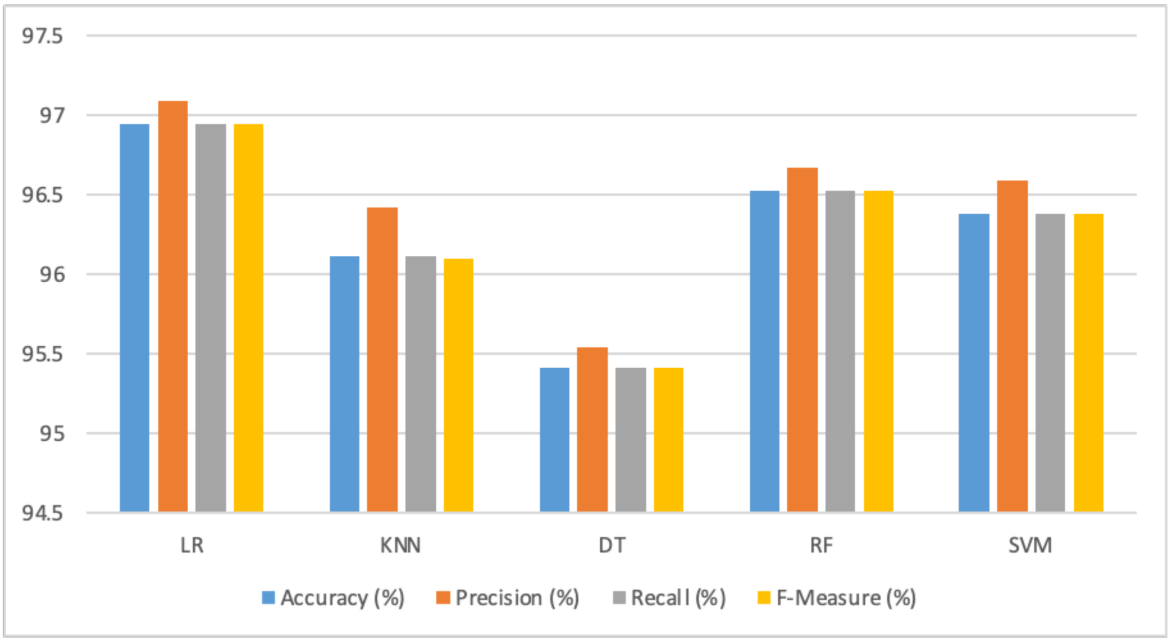

Figure 6. Describes the different stacking models with data balancing, FS (chi^2) & without hyperparameter tuning.

4.5 Generation results were obtained using a 10-fold CV with Data balancing, FS (Correlation) & without hyperparameter tuning

The empirical results obtained for the WBCD dataset are shown in Table 10. Performance-wise, the stacking model SVM and DT, LR, RF classifiers as estimators performed better than other ways in the 10-fold CV.The table delineates the performance of several models inside a stacking framework, evaluated using four metrics: accuracy, precision, recall, and F-measure. The stacking model with DT, KNN, LR, RF, and SVM as base estimator achieved the highest accuracy (97.64%), precision (97.78%), recall (97.64%), and F-Measure (97.63%) among the stacking framework. Figure 7 visually illustrates the accuracy of different stacking models with data balancing, FS (Correlation) & without hyperparameter tuning.

Table 10: Describe the results that Data balancing and FS (Correlation) & without hyperparameter tuning

Estimators	Stacking Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
DT, SVM, KNN, and RF	LR	96.94	97.06	96.94	96.94
DT, SVM, RF, and LR	KNN	96.80	96.97	96.80	96.79
SVM, RF, LR and KNN	DT	94.99	95.16	94.99	94.98
DT, KNN, SVM, and LR	RF	96.24	96.48	96.24	96.24
DT, KNN, LR, and RF	SVM	97.64	97.78	97.64	97.63

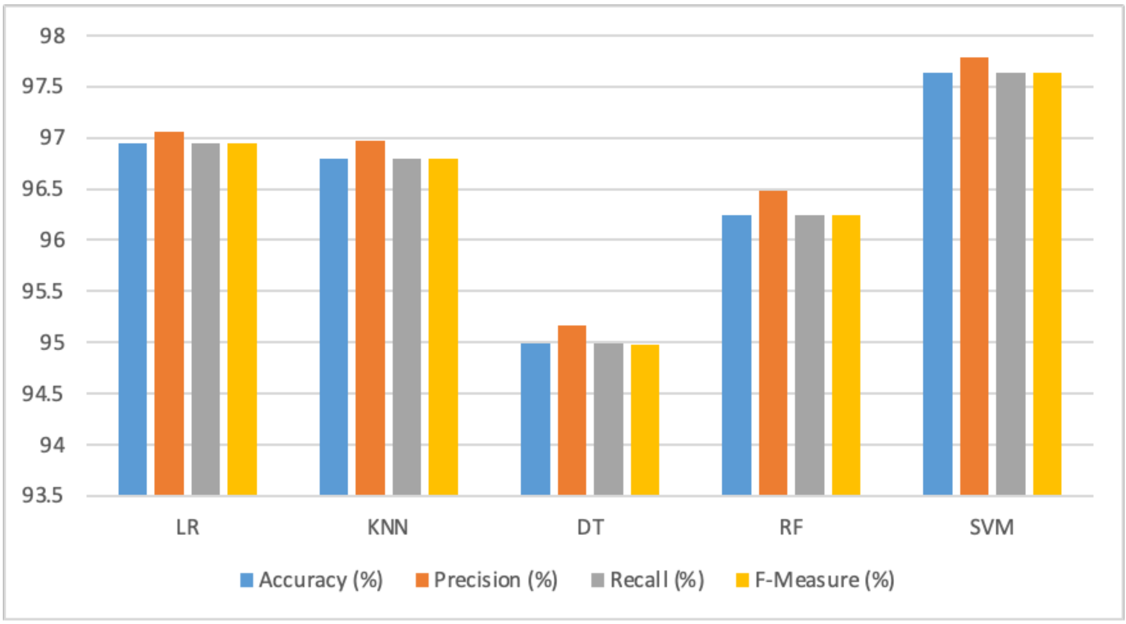

Figure 7. Describes the different stacking models with data balancing, FS (Correlation) & without hyperparameter tuning

4.6 Comparative analysis of previous work

The associated work's comparative summary is shown in Table 11. The table compares the best accuracy achieved by various studies using the WBCD dataset for BC prediction. Notable results include 94.74% by Rahmanul Hoque et al. (2024), 96.44% by Rahul Karmakar et al. (2023), 97.40% by Hua Chen et al. (2023), 97.00% by Varsha Nemade et al. (2023), 96.50% by Annisa Maulidia et al. (2023), 95.75% by Saheb Karan et al. (2023), 95.00% by Amit Kumar Jakhar et al. (2023), 97.37% by Sara Laghmati et al. (2023), and 97.36% by Hajar Saoud et al. (2019). The proposed work outperformed all previous studies with the highest accuracy of 97.63%, demonstrating incremental improvements in prediction accuracy over time. The proposed work tackles the issue of data imbalance, an aspect often overlooked in previous studies. While oversampling can help address this problem, it introduces noise and increases the risk of overfitting, ultimately compromising the model's performance. By contrast, the use of 10-fold cross-validation (CV) provides a more accurate variance estimate for the model. This is achieved through iterative testing and training, which reduces the impact of random sampling and delivers a more robust evaluation of the model's performance.

Table 11: Comparison with the previous works

Reference	Dataset	Accuracy (Best in %)
Rahmanul Hoque et al. (2024)[10]	WBCD	94.74
Rahul Karmakar et al.(2023)[11]	WBCD	96.44
Hua Chen et al. (2023)[12]	WBCD	97.40
Varsha Nemade et al. (2023) [13]	WBCD	97.00
Annisa Maulidia et al. (2023) [14]	WBCD	96.50
Saheb Karan et al. (2023) [15]	WBCD	95.75
Amit Kumar Jakhar et al. (2023) [16]	WBCD	95.00
Sara Laghmati et al.(2023)[18]	WBCD	97.37
Hajar Saoud et al.(2019)[22]	WBCD	97.36
Proposed work	WBCD	**97.63**

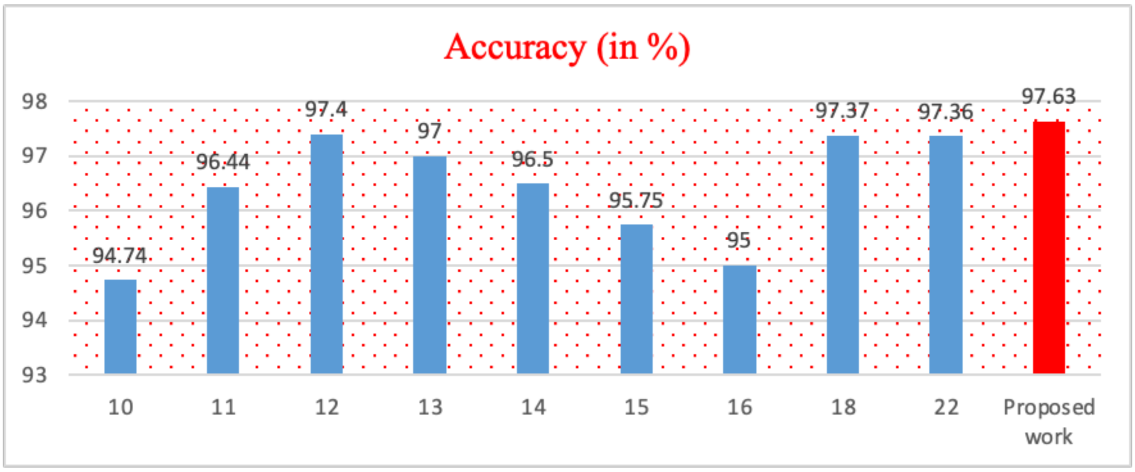

Figure 8. Describe the different stacking models with data balancing, FS (Correlation) & without hyperparameter tuning

5 Conclusion and future work

Many pieces of equipment used in the healthcare sector nowadays generate enormous volumes of data. Predicting diseases requires the identification and analysis of this data, which is a major topic of research. ML evolution is a savior for predictive analysis of such huge datasets. In order to identify and assess BC, our study employed five distinct stacking model forecasting techniques to estimate the BC. LR was used as the stacking classifier in the beginning, and the estimators were DT, SVM, KNN, and RF classifiers. Then, we chose SVM, KNN, RF, and LR as estimators and DTs as the stacking model. We used DTs, LR, KNN, and RF as estimators and SVM as a stacking classifier, LR, DT, KNN, and RF as estimators and KNN as a stacking model, following that, LR, DT, and KNN, SVM as estimators and RF as a stacking classifier.

The results of this study demonstrate the effectiveness of the proposed stacking model for breast cancer prediction using the WBCD dataset. Several experiments were conducted with different configurations, including varying combinations of data balancing, FS, and hyperparameter tuning. The analysis showed that the best performance, with an accuracy of 97.63%, was achieved when data balancing and hyperparameter tuning were applied without FS. Without FS, Hyperparameter Tuning, or Data Balancing, the stacking model achieved 95.96% accuracy, providing a baseline performance. With Data Balancing but Without FS or Hyperparameter Tuning, accuracy improved to 96.94%, showing the positive impact of data balancing. With Data Balancing and Hyperparameter Tuning, the model further improved to 97.63% accuracy, emphasizing the importance of tuning for better performance. Feature Selection (Chi^2 or Correlation), when combined with data balancing, helped enhance precision and recall, although the best performance was observed without FS in certain configurations. Compared to previous works, the proposed model outperformed all other studies on the WBCD dataset, achieving the highest accuracy of 97.63%. This reflects incremental advancements in prediction accuracy over time, demonstrating the robustness of the stacking model in addressing class imbalances and improving prediction outcomes for breast cancer detection.

Future studies could explore the impact of additional feature selection techniques, alternative meta-learners, and deep learning methods to further improve prediction accuracy. Additionally, integrating more diverse datasets and experimenting with larger ensemble models could enhance the model's generalization ability. Finally, incorporating real-time data and further optimizing the model for deployment in clinical settings may facilitate more practical and timely breast cancer detection.

References

[1] Prusty S, Patnaik S, Dash SK, Prusty SG. SEMeL-LR: An improvised modeling approach using a meta-learning algorithm to classify breast cancer. Engineering Applications of Artificial Intelligence. 2024 Mar 1;129:107630. doi: 10.1016/j.engappai.2023.107630.

[2] Consoli S, Recupero DR, Petkovic M. Data Science for Healthcare: Methodologies and Applications. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-05249-2
[3] Kadhim RR, Kamil MY. Comparison of machine learning models for breast cancer diagnosis. IAES International Journal of Artificial Intelligence. 2023 Mar 1;12(1):415. doi: 10.11591/ijai.v12.i1.pp415-421
[4] Alnowami MR, Abolaban FA, Taha E. A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer. Journal of Radiation Research and Applied Sciences. 2022 Mar 1;15(1):104-10. doi: 10.1016/j.jrras.2022.01.003
[5] Das AK, Biswas SK, Mandal A, Bhattacharya A, Sanyal S. Machine learning based intelligent system for breast cancer prediction (MLISBCP). Expert Systems with Applications. 2024 May 15;242:122673. https://doi.org/10.1016/j.eswa.2023.122673
[6] Das AK, Biswas SK, Mandal A. An expert system for breast cancer prediction (ESBCP) using decision tree. Indian J Sci Technol. 2022 Feb 12;15(45):2441-50. https://doi.org/ 10.17485/IJST/v15i45.756
[7] Hussain A, Xu PC, Shixin W, Qureshi KN. Artificial Intelligence in Natural Science Research: Innovations and Limitations. InNext Generation AI Language Models in Research 2025 (pp. 129-169). CRC Press.DOI: 10.1201/9781032667911-6
[8] Aslam A, Qureshi KN, Hussain A. Advanced AI-Based Healthcare Systems Applications and Services. InNext Generation AI Language Models in Research 2024 Nov 13 (pp. 86-113). CRC Press. https://doi.org/10.1201/9781032667911
[9] Prusty S, Patnaik S, Dash SK, Prusty SG. SEMeL-LR: An improvised modeling approach using a meta-learning algorithm to classify breast cancer. Engineering Applications of Artificial Intelligence. 2024 Mar 1;129:107630.doi: 10.1016/j.engappai.2023.107630.
[10] Hoque R, Das S, Hoque M, Haque E. Breast Cancer Classification using XGBoost. World Journal of Advanced Research and Reviews. 2024;21(2):1985-94.doi: 10.30574/wjarr.2024.21.2.0625
[11] Karmakar R, Chatterjee S, Das AK, Mandal A. BCPUML: Breast Cancer Prediction Using Machine Learning Approach—A Performance Analysis. SN Computer Science. 2023 May 6;4(4):377. https://doi.org/10.1007/s42979-023-01825-x
[12] Chen H, Wang N, Du X, Mei K, Zhou Y, Cai G. Classification prediction of breast cancer based on machine learning. Computational intelligence and neuroscience. 2023;2023(1):6530719. https://doi.org/10.1155/2023/6530719
[13] Nemade V, Fegade V. Machine Learning Techniques for Breast Cancer Prediction. Procedia Computer Science. 2023 Jan 1;218:1314-20. https://doi.org/10.1016/j.procs.2023.01.110
[14] Maulidia A, Lidyawati L, Jambola L, Kristiana L. Analysis of logistic regression algorithm for predicting types of breast cancer based on machine learning. InAIP Conference Proceedings 2023 Feb 24 (Vol. 2772, No. 1, p. 040005). AIP Publishing LLC. https://doi.org/10.1063/5.0135393
[15] Karan S,Chowdhury AR, Pal A, Das S, Das S, Chaudhuri AK. Early Detection Of Breast Cancer Using Logistic Regression Method. International Journal of Engineering Technology and Management Sciences. 2023;7(2):133-42. https://doi:10.46647/ijetms.2023.v07i02.017
[16] Jakhar AK, Gupta A, Singh M. SELF: a stacked-based ensemble learning framework for breast cancer classification. Evolutionary Intelligence. 2023 Jan 31:1-6. https://doi.org/10.1007/s12065-023-00824-4.
[17]Patel R.Diagnosis of Breast Cancer using Data Mining Techniques. Journal For Basic Sciences.2023;23(4), DOI: 10.37896/JBSV23.4/2007.
[18]Laghmati S, Hamida S, Hicham K, Cherradi B, Tmiri A. An improved breast cancer disease prediction system using ML and PCA. Multimedia Tools and Applications. 2024 Mar;83(11):33785-821. doi: 10.1007/s11042-023-16874-w.
[19] Rahman MM, Ferdousi Z, Saha P, Mayuri RA. A machine learning approach to predict breast cancer using boosting classifiers. Indian J. Comput. Sci. Eng. 2023 Jun;14(3):409-15. doi: 10.21817/indjcse/2023/v14i3/231403009.
[20] Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science. 2021 Jan 1;191:487-92. doi: 10.1016/j.procs.2021.07.062

[21] Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, Podder E. Breast cancer risk prediction using XGBoost and random forest algorithm. In2020 11th international conference on computing, communication and networking technologies (ICCCNT) 2020 Jul 1 (pp. 1-4). IEEE. doi: 10.1109/ICCCNT49239.2020.9225451.
[22] Saoud H, Ghadi A, Ghailani M, Abdelhakim BA. Using feature selection techniques to improve the accuracy of breast cancer classification. InInnovations in Smart Cities Applications Edition 2: The Proceedings of the Third

International Conference on Smart City Applications 2019 (pp. 307-315). Springer International Publishing. doi: 10.1007/978-3-030-11196-0_28.
[23] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic[Online acess:12-5-2024]
[24] http://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original[Online acess:12-5-2024]