A Cascading Ensemble with Custom Subset Generation and Multi-Level Fusion for Enhanced Breast Cancer Detection

Vandana Lingampally[1], Dr. K. Radhika[2]

[1]Research scholar, Department of CSE, Osmania University, Hyderabad, Telangana, India.
[2]Professor, Department of IT, CBIT, Hyderabad, Telangana, India

Abstract

The ensemble learning technique is a powerful method that combines multiple machine learning models to address their individual limitations and create an optimized predictive model. By harnessing the strengths of each model, ensemble learning significantly enhances overall performance, making it particularly valuable for classification tasks across diverse domains. In this study, an efficient three-level stacking ensemble approach is proposed for diagnosing breast cancer. At the first level, a diverse set of base learners is employed, encompassing decision trees, logistic regression, k-nearest neighbors, support vector machines, and Gaussian naive bayes. Each first-level learner is trained on distinct subsets of the training data with 10-fold cross-validation, thereby ensuring model robustness and mitigating the risk of overfitting. Transitioning to the second level, sophisticated ensemble models including Adaboost, GBM, and Random forest are introduced. These second-level learners undergo training using the validation predictions generated by the first-level models, while incorporating the top six informative features extracted from the dataset. To enhance their performance and mitigate overfitting, the models are optimized through the application of 10-fold cross-validation. In order to achieve further refinement of the ensemble, a third-level learner, represented by a deep neural network (DNN), is introduced. This DNN is trained on the validation predictions obtained from both the first and second levels, facilitating the capturing and synthesis of the collective knowledge of the ensemble. By leveraging the inherent capabilities of deep learning, the DNN maximizes the amalgamated benefits derived from the preceding levels, resulting in a substantial enhancement of the overall predictive power. The effectiveness of the ensemble approach is evaluated using well-established performance metrics, including accuracy, sensitivity, and specificity, which are derived from the analysis of the confusion matrix. The results incontrovertibly demonstrate that the ensemble models surpass individual machine learning models in accurately detecting breast cancer.

Keywords: breast cancer, machine learning, ensemble techniques, stacking.

1. INTRODUCTION

Why Ensemble?

Ensemble learning is a powerful technique that combines multiple machine learning models to address the limitations of individual models and improve predictive performance. It leverages the diversity of these models, incorporating different algorithms, feature representations, and subsets of training data. By doing so, ensemble learning mitigates biases and limitations, capitalizes on strengths, and enhances overall accuracy, robustness, and generalization. Ensemble reduces bias, variance, and overfitting, capturing the collective knowledge of constituent models for superior predictive power, especially in complex datasets. They also provide robustness against noise and outliers, ensuring reliability in challenging environments. Additionally, ensemble methods enhance interpretability by offering insights into underlying patterns, feature importance, and decision-making processes, facilitating better understanding and integration of domain knowledge.

Ensemble Techniques

There are various types of ensembles in ensemble learning, each with its own characteristics and strengths. Few of them are

Bagging: Bagging, also known as Bootstrap Aggregating is a widely used ensemble method that involves training multiple base models on distinct bootstrap samples taken from the original training data. These models' predictions are then combined, typically through voting or averaging, to generate the final ensemble prediction. By incorporating various perspectives from different subsets of the data, bagging effectively diminishes variance and enhances stability.

Boosting: Boosting is another powerful ensemble method that sequentially builds a strong model by focusing on data points that were previously misclassified. The base models in boosting are trained iteratively, with each subsequent model attempting to correct the errors made by the previous models. Boosting algorithms, such as Adaboost, Gradient Boosting, and XGBoost, can effectively handle complex relationships in the data and achieve high predictive accuracy.

Stacking: Stacking, also known as stacked generalization, involves training multiple diverse base modelson the same dataset. However, instead of aggregating their predictions directly, a meta-model (or blender) is trained to learn how to combine the predictions of the base models. Stacking allows for more sophisticated modeling and can capture higher-order relationships between the base models outputs.

Stacking Model

The stacking ensemble method, also known as stacked generalization, is a powerful technique that combines the predictions of multiple diverse base models through a meta-model or blender. Unlike other ensemble methods, stacking goes beyond simple aggregation and learns to combine the base models predictions effectively. It exploits the strengths and diversity of the base models by training a meta-model to capture a broader range of information and potentially improve overall performance and generalization. The stacking process involves a training phase where the base models are trained on the same dataset, and their predictions are collected as new features. A meta-model is then trained to weigh

the base models predictions and produce the final prediction. In the prediction phase, the base models predictions are fed into the trained meta-model to generate the final prediction.

The stacking ensemble method offers several advantages. Firstly, it can capture complex relationships and interactions between the predictions of the base models, allowing for more sophisticated modeling. Secondly, it can learn to correct the biases or weaknesses of individual base models, leading to improved predictive accuracy. Lastly, stacking provides insights into the importance and relevance of the base models, as well as the relationships between them. By intelligently combining the predictions of diverse models, stacking enables a more robust and powerful ensemble approach in machine learning.

Three Level Stacking Model

In this present research, a three-level stacking model was developed that incorporates distinct classifiers at each level. At the first level, five diverse classifiers were we employed: decision trees, logistic regression, k-nearest neighbors, support vector machines, and Gaussian naive bayes. This selection aims to capture a wide range of modeling approaches and account for different data characteristics. Moving to the second level, three classifiers are utilized: Adaboost, Gradient Boosting Machine, and Random Forest. These classifiers are known for their robustness and ability to handle complex patterns in the data. Finally, at the third level, we incorporated a deep neural network classifier. Deep neural networks are renowned for their capacity to learn intricate representations and capture high-level features in the data. By employing this hierarchy of classifiers with varying strengths and capabilities, the three-level stacking model aims to leverage the collective wisdom and diversity of the classifiers to enhance overall prediction performance.

Short note of Classifiers

Decision Trees

Decision trees are versatile and interpretable machine learning models widely used for classification and regression tasks. They offer clarity and understanding in the decision-making process by providing transparency through explicit feature and threshold associations for each split. This attribute facilitates easy interpretation of the model's logic. They possess the flexibility to handle both numerical and categorical data, allowing their application to diverse datasets. Moreover, they excel in capturing intricate non-linear correlations between features and the target variable, resulting in a resilient and versatile model capable of adapting to different data patterns. Additionally, decision trees can handle missing values and outliers, and they are not sensitive to feature scaling.

Logistic Regression

Logistic regression is a commonly employed statistical learning technique utilized for binary classification tasks. Its purpose is to establish a model that relates the input variables to the probability of the target variable belonging to a specific class. Through fitting a logistic function to the data, logistic regression estimates coefficients that gauge the impact of each input variable on the predicted outcome. One of its key advantages is interpretability, as these coefficients represent the log-odds of the target variable, providing valuable insights into the variables importance and direction. Logistic regression is computationally efficient, exhibits good performance with large datasets, and is capable of handling both

continuous and categorical input features. Its widespread usage spans various domains such as finance, healthcare and social sciences, where comprehending the likelihood of binary outcomes is crucial.

K-Nearest Neighbours

K-nearest neighbors (KNN) is a versatile machine learning algorithm utilized for both classification and regression tasks, operating on the principle that similar data points often share the same class or exhibit similar output values. By considering the K nearest neighbors from the training set, KNN determines the predicted class or value for a given data point. The algorithm calculates the distance between data points employing different distance metrics like Euclidean or Manhattan distance. KNN stands out for its simplicity and intuitive nature since it doesn't require any assumptions regarding the underlying data distribution. It is applicable to a wide range of data types, including numerical and categorical. However,the choice of K and the distance metric can impact KNN's performance, and it may become computationally intensive for large datasets. Nevertheless, KNN finds extensive application in diverse domains such as pattern recognition, recommender systems, and anomaly detection.

Support Vector Machine

Support Vector Machine (SVM) is powerful machine learning algorithm used for classification and regression tasks. They aim to find the best decision boundary that maximally separates different classes by transforming data into a higher-dimensional feature space. SVMs can handle nonlinear data through the kernel trick, which implicitly maps data into higher dimensions. SVMs are efficient with high-dimensional data and have a lower risk of overfitting. They have been successfully used in various domains such as text categorization and image recognition. SVMs suitable for scenarios where the available data is limited Support vectors are the data points that lie closest to the decision boundary and they play a crucial role in defining the decision boundary and the overall SVM model.

Gaussian Naive Bayes

Gaussian Naive Bayes is a reliable and straightforward machine learning algorithm employed for classification tasks, leveraging Bayes' theorem. Its core assumption is that the features are conditionally independent given the class variable. By modeling the likelihood of features using a Gaussian distribution, it is well-suited for continuous input variables. It evaluates the posterior probability of each class based on the input features and assigns the data point to the class with the highest probability. Although it assumes feature independence, it demonstrates effective performance in numerous real-world applications. It boasts computational efficiency, requires minimal training data, and remains robust against irrelevant features. The algorithm finds successful utilization across various domains, including text classification, spam filtering, and sentiment analysis, where the assumption of feature independence holds reasonably well.

AdaBoost

Adaboost, also known as Adaptive Boosting, is an ensemble technique that enhances the performance of weak classifiers by combining them into a powerful classifier. It accomplishes this by assigning higher weights to samples that were previously misclassified, enabling subsequent weak classifiers to focus on those challenging instances. Through iterative updates, Adaboost adapts to the data by adjusting the weights, thereby improving the overall accuracy by leveraging the collective knowledge of the weak classifiers.By iteratively focusing on misclassified samples, Adaboost pays more attention to

difficult instances and gradually reduces the errors made by the ensemble model. This adaptive nature helps prevent overfitting, making Adaboost a robust algorithm that generalizes well to unseen data.

Random Forest

Random Forest is a flexible ensemble algorithm that harnesses the power of multiple decision trees to construct a resilient and precise model. By training each decision tree on a random subset of features and data samples, the ensemble achieves diversity. The algorithm combines the predictions of individual trees, typically through voting or averaging, to generate final predictions. Random Forest stands out for its capacity to effectively handle high-dimensional data, capture intricate interactions, and address missing values. It showcases robustness against overfitting and delivers excellent performance in both classification and regression tasks.

Gradient Boosting Machine

Gradient Boosting Machine is a powerful boosting algorithm widely used in machine learning. It sequentially builds a series of weak learners, typically decision trees, by iteratively learning from the mistakes made by the previous models. Each subsequent tree focuses on reducing the errors of the ensemble, aiming to improve overall predictive accuracy. GBM is particularly effective in capturing complex non-linear relationships, and achieving high performance across various domains.

Deep Neural Network

Unlike traditional neural networks with only a few hidden layers, DNNs leverage their depth to capture intricate hierarchical features in the data. Each layer extracts progressively abstract representations from the input data, allowing the network to learn high-level representations that lead to accurate predictions. This depth enables DNNs to handle large-scale datasets and complex problem domains effectively. DNNs offer a notable benefit by autonomously acquiring features from the data, diminishing the necessity for manual feature engineering. Consequently, they possess a remarkable adaptability to various tasks and domains.

Ensemble learning technique, the three-level stacking ensemble approach described in this study, offers valuable insights and advancements in the field of machine learning. By combining the strengths of different classifiers in a hierarchical structure, ensemble models enhance predictive power, robustness, and generalization capabilities. The use of diverse classifiers at each level enriches the ensemble's ability to capture complex data patterns.

This paper follows a structured approach to present the research findings. Firstly, a comprehensive review of relevant prior studies in the field was conducted (Section 2). Subsequently, detailed explanation of the techniques employed in our proposed model is provided(Section 3). The implementation process is outlined, and the results of the study are presented and analyzed in depth (Section 4). Lastly, the key findings obtained from the research and propose potential directions for future investigations is summarized(Section 5).

2. LITERATURE REVIEW

Breast cancer poses a substantial global health challenge, underscoring the critical importance of early detection in optimizing patient prognosis. To this end, researchers have embarked on harnessing the potential of machine learning, particularly ensemble models, to bolster the accuracy and dependability of breast cancer detection. In the realm of scientific inquiry, a multitude of studies have emerged, introducing and assessing diverse ensemble models tailored to breast cancer classification across a wide spectrum of datasets and machine learning methodologies. By delving into this body of literature, this review aims to shed light on the various ensemble models proposed and their corresponding evaluations, ultimately contributing to the advancement of breast cancer detection practices.

In their study [1], the authors proposed four ensemble models, namely NFE, KNNE, QCE, and a combination of NF+KNN+QC, for the detection of breast cancer (BC). Through their experiments using the WBCD dataset, they demonstrated that the ensemble model NF+KNN+QC outperformed the other models in terms of BC detection, achieving an accuracy of 97.14%. The individual models, NFE, KNNE, and QCE, achieved accuracies of 96.56%, 96.42%, and 96.57%, respectively. In their work [2], the author conducted a comprehensive comparison of the performance of six ensemble methods, namely Bagging, Dagging, Ada Boost, Multi Boost, Decorate, and Random Subspace, using the WBCD dataset for breast cancer detection. These ensemble methods were evaluated in conjunction with fourteen base learners, including Bayes Net, FURIA, KNN, C4.5, RIPPER, KLR, K-star, LR, MLP, Naive Bayes, RF, Simple Cart, SVM, and LMT. The results revealed that the highest accuracy of 97.66% was achieved by the Bayes Net algorithm in combination with the dagging ensemble method. On the other hand, the Single RIPPER algorithm exhibited the lowest accuracy of 95.17% among all the base learners evaluated. This study highlights the significance of employing ensemble methods to improve the accuracy of breast cancer detection.

In their research paper [3], the authors devised an ensemble model that leveraged the capabilities of five distinct base learning algorithms, namely Naive Bayes, Decision Tree with Gini index, Decision Tree with Information Gain, Support Vector Machine, and MBL (a specified algorithm). The ensemble model employed a weighted voting scheme to consolidate the predictions generated by these diverse algorithms. Through their experimentation, the ensemble model achieved an impressive accuracy of 97.42% in the task at hand. In the research paper [4], a novel approach for breast cancer detection was introduced, which involved the utilization of an ensemble consisting of neural networks. This ensemble-based model exhibited remarkable performance, achieving an accuracy of 96.43% when evaluated on the widely-used WBCD dataset. The paper [5] presents a novel approach for breast cancer detection by proposing an ensemble model based on Support Vector Machine. The authors designed the ensemble by incorporating twelve base learners, comprising six different SVM kernels with both C-SVM and V-SVM variations. The model was trained and evaluated on multiple datasets, including the WBCD (original and diagnostic) and SEER datasets. Through their comprehensive experiments, the authors achieved a remarkable accuracy of 97.68% using their SVM-based ensemble model.

The research paper [6] introduces a stacking ensemble technique for breast cancer detection. The proposed model follows a two-step classification process, incorporating both base learners and meta learners. Specifically, the base learners utilized in the model include GBM, DRF (Distributed Random Forest), GLM (Generalized Linear Model), and DNN. The performance of this model was evaluated using two datasets: WBCD (original and diagnostic). Through extensive experimentation, the authors achieved remarkable results, with the highest accuracy of 97.96% obtained by the meta learner GBM. In their study [7], the authors proposed an innovative approach that builds upon existing ensemble models for breast cancer detection by incorporating feature selection techniques. Specifically, F-test and variance thresholding methods were employed to identify the most significant features within the dataset. The ensemble model consisted of bagging and boosting techniques using base learners such as SVM with an RBF (Radial Basis Function) kernel, Naive Bayes, and SVM with linear kernel. The meta learner for this model was logistic regression. The performance of the model was evaluated using the WBCD (original and diagnostic) datasets, as well as the microRNA dataset. The results demonstrated the model's effectiveness, achieving a maximum accuracy of 97.37%.

In their study [8], the authors introduced four ensemble models with three base learners and two or three meta learners. The base learners employed were BayesNet, LMT, and SGD for the BayesNet-2/3 MetaClassifier model, and Naive Bayes, LMT, and SGD for the Naive Bayes-2/3-MetaClassifier model. The meta learners utilized for the two-meta classifier models were SGD and J48, while for the three-meta classifier models, SGD, J48, and REPTree were used as meta learners. These ensemble models were evaluated on the WBCD dataset using k-fold cross-validation, with values of k set as 3, 5, and 10. Remarkably, the highest accuracy of 98.07% was achieved by the 2/3-meta classifier models when the value of k was set to 10. In their study [9], the authors introduced a stacking ensemble model based on logistic regression for breast cancer detection. The model was meticulously evaluated on the WBCD dataset using a 3-fold cross-validation methodology. Impressively, the LR stacking ensemble achieved a remarkable accuracy of 98.96%.

This paper [10] introduces SELF, a stacked-based ensemble learning framework for early-stage breast cancer classification using histopathological images. The BreakHis dataset with 7909 images and the WBCD with 569 instances were used for evaluation. The authors trained multiple classifiers and selected the top five based on accuracy for their ensemble model. The chosen classifiers were Extra tree, Random Forest, Adaboost, Gradient Boosting, and KNN9, with a logistic regression model as the final estimator. SELF achieved testing accuracies of approximately 95% and 99% on the BreakHis and WBCD datasets, respectively. The framework also demonstrated superior performance in terms of F1-Score, ROC, and MCC scores on the BreakHis dataset. This paper [11] the authors present a novel approach for classifying Breast Cancer as benign or malignant using feature ensemble learning based on Sparse Autoencoders and Softmax Regression. The study utilized the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI machine learning repository. Performance evaluation included multiple indices such as accuracy, specificity, sensitivity, recall, precision, F-measure, and MCC. The proposed approach demonstrated superior results across various parameters, achieving a true accuracy of 98.60%.

This study [12] introduces a combined approach of ensemble methods and imbalanced learning techniques for breast cancer classification. The methodology involves applying the Synthetic Minority Over-Sampling Technique (SMOTE), an imbalanced learning algorithm, to the selected datasets. Additionally, Bayesian Optimization is employed to fine-tune multiple baseline classifiers. The proposed ensemble model, incorporating SMOTE, demonstrates high accuracy in classifying instances as either benign or malignant. The approach achieves an overall accuracy of 98.14% on the WBCD Original dataset and 97.45% on the Diagnostic dataset. Previous research in ensemble learning has predominantly concentrated on ensembles consisting of one or two levels, where the classifiers employed within a level or two levels tend to be similar. However, this focus has resulted in a dearth of experimental analysis regarding the combination of classifiers, the optimal number of levels, and their collective impact on ensemble performance. To address this research gap comprehensively, An innovative three-level stacking ensemble that incorporates diverse classifiers across different levels is proposed. Moreover, additional strategies to enhance the robustness and efficacy of the model were introduced. By undertaking these advancements, a more refined and sophisticated approach to ensemble designs is achieved, thereby contributing to the progress of ensemble learning.

3. PROPOSED APPROACH

Dataset

The WBCD dataset is utilized for evaluating the effectiveness of the proposed model. This dataset was collected from patients who received treatment at the hospitals affiliated with the University of Wisconsin-Madison. It includes 699 samples. Every sample has nine features. The dataset includes cytological features of breast fine needle aspiration test results, where each feature is assigned a value between one and ten, with higher values indicating greater intensity. The final attribute, Class, can be either benign with a value of 2 or malignant with a value of 4. Within the dataset, there are 458 instances classified as benign cases and 241 instances classified as malignant cases. 16 samples have missing values in one attribute bare nuclei, these are replaced by mean value during data pre-processing. Table 1 shows the attributes of the dataset.

FusionCascade: A Cascading Ensemble with Custom Subset Generation and Multi-Level Fusion(CECSMF)

The proposed model utilizes sophisticated 3-level stacking ensemble architecture, designed to enhance predictive performance and exploit the complementary strengths of multiple classifiers.

1. The first level of the ensemble comprises a diverse set of five classifiers: Decision Trees , Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Gaussian Naive Bayes. These classifiers serve as the foundation for the subsequent levels of the ensemble.

2. To foster diversity and robustness, the training dataset is partitioned into five subsets through a random sampling scheme, ensuring each subset shares the same size as the original training dataset. This controlled random sampling process adheres to specific constraints aimed at capturing different perspectives of the data.

3. Each classifier in the first level is trained on a distinct subset using a rigorous 10-fold cross-validation strategy. This technique aids in assessing the classifiers' performance by iteratively partitioning the data into training and validation sets, leading to more reliable estimates of their

effectiveness. The trained classifiers are then evaluated using the independent test dataset.

4. The predictions generated by the first level classifiers on both the validation and test datasets are forwarded to the subsequent level of the ensemble. This process enables the stacking ensemble to capitalize on the collective insights derived from the first level classifiers.

5. The second level of the ensemble comprises three high-performing classifiers: Adaboost, Gradient Boosting Machine and Random Forest. These classifiers are trained on the validation predictions obtained from the first level classifiers. Furthermore, the top six most informative features, selected based on the ranking given by the RF, are incorporated into the training process. The evaluation of these classifiers is conducted using the test predictions derived from the first level combined with top 6 features.

6. The training procedure for the second level classifiers involves employing a robust 10-fold cross-validation approach. This approach ensures an accurate estimation of their predictive capabilities while leveraging the fused predictions and relevant features from the first level.

7. The third and final level of the ensemble is implemented using a Deep Neural Network (DNN). This neural network model is trained on the validation predictions obtained from both the first and second levels. By combining the outputs of the preceding levels, the DNN can effectively capture and learn complex patterns and dependencies within the data. The evaluation of the DNN's performance is assessed with the test predictions generated at both the first and second levels.

8. The predictions produced by the DNN at the third level serve as the ultimate prediction outcome of the stacking ensemble. This final prediction reflects the amalgamation of knowledge and insights acquired from multiple levels and classifiers within the ensemble, offering enhanced predictive power and robustness. Fig.1 presents flow chart of proposed model, Fig.2 architecture of the proposed model and Fig.3 presents diagrammatical presentation of proposed model.
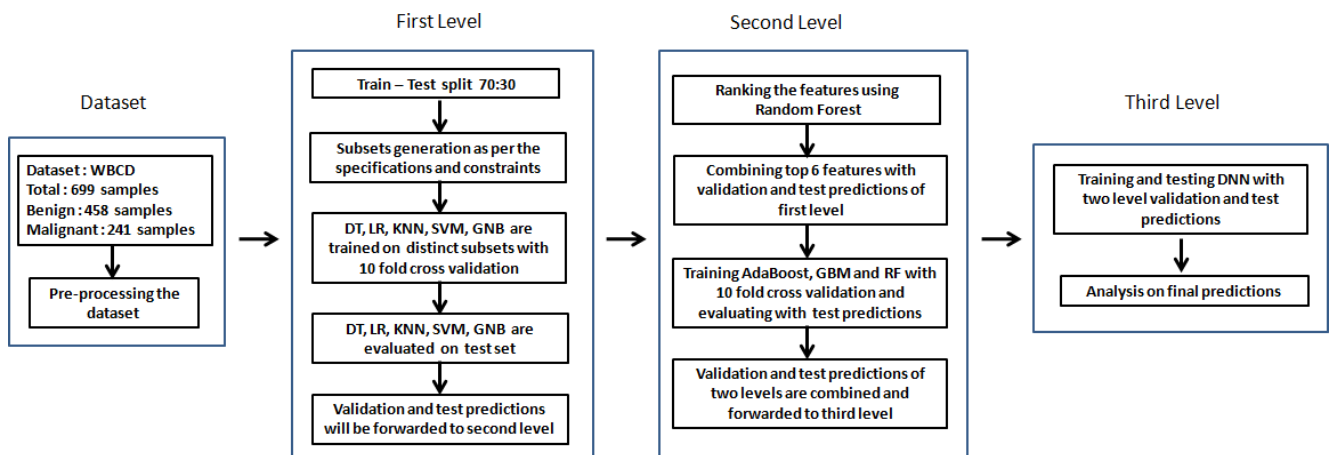

Fig.1. Flow chart of the CECSMF model

Subset Generation

The process of generating the subsets is as follows

1. Determine the desired size of each subset based on the training set size.
2. Calculate the number of benign and malignant samples to be allocated to each subset. Divide the respective counts by 2.
3. Create five empty subsets: S1, S2, S3, S4, and S5.
4. Allocate 50% of the benign samples to all subsets, ensuring an equal distribution across subsets.
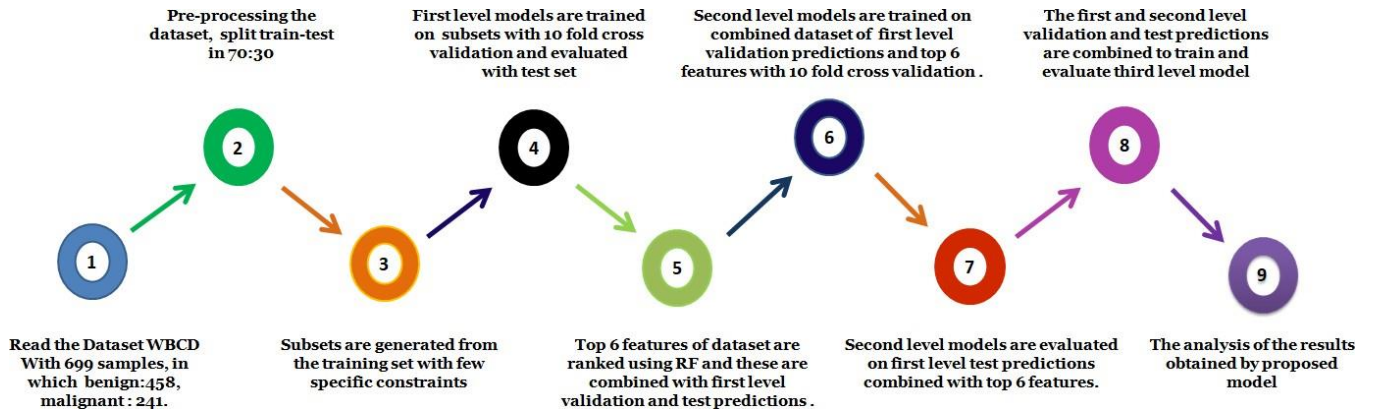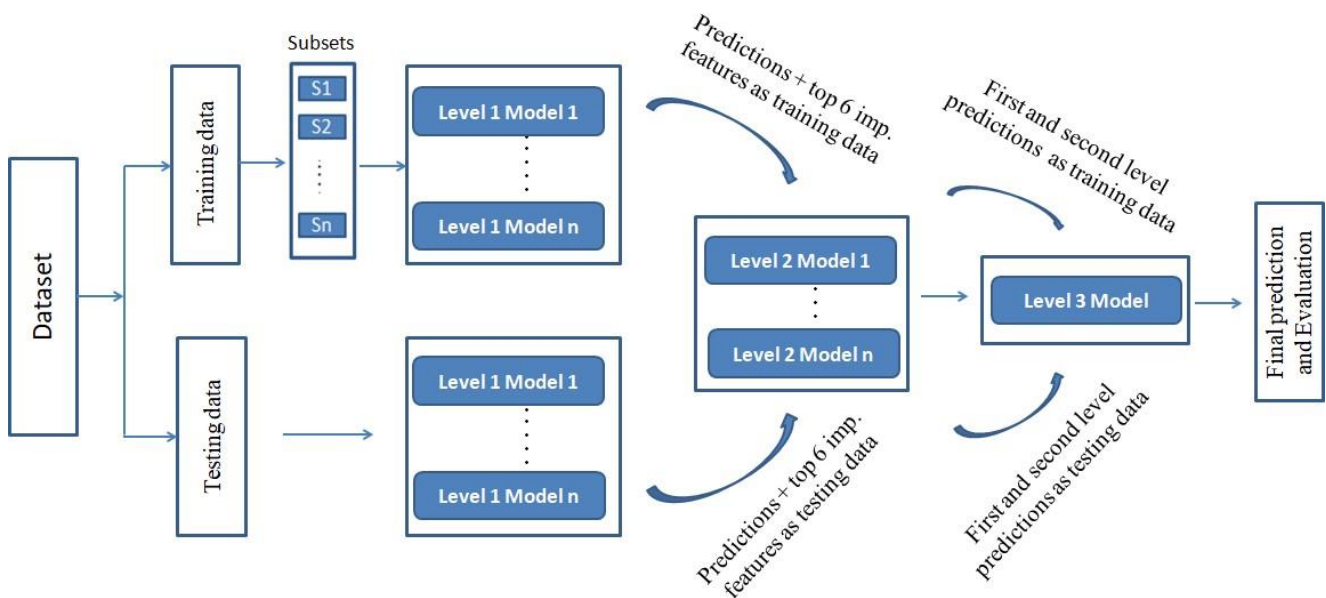
Fig.2. Architecture of the CECSMF model

Fig.3. Diagrammatical presentation of CECSMF model

5. Allocate 50% of the malignant samples to all subsets, ensuring an equal distribution across subsets.

6. Repeat the following steps until the size of each subset matches the desired training set size, the number of benign and malignant samples is equal proportion in each subset, and no sample appears more than three times in any subset:
   a. Randomly select 50 samples, with replacement, from the remaining benign and malignant samples, ensuring a mix of benign and malignant samples based on the requirement.
   b. Add the selected samples to each subset, ensuring an equal distribution across subsets.

7. Before the last random sampling iteration, check for any samples that has not been allocated to any subset. Use those samples in the next allocation.

8. Once the size of each subset matches the desired training set size, the number of benign and malignant samples is in same proportion in each subset, and no sample appears more than three times in any subset, the generation process is complete.

Combining the validation predictions

The initial-level models are trained on subsets generated by randomly selecting samples from the dataset, which allows them to capture a wider range of diverse patterns within the data. These models generate validation predictions, which are then used in the second level to train classifiers. As a result of the random selection process, some samples may appear multiple times in different subsets and have corresponding predictions. To address this, a maximum voting scheme is employed to combine the predictions for these samples. This scheme selects the class label that receives the highest number of votes, thereby aggregating the predictions across the subsets. By utilizing this voting mechanism, the combined predictions effectively account for the repeated samples, ensuring their contributions are appropriately considered in the overall decision-making process.

Selection of top six features

Feature selection plays a crucial role in data analysis and ML tasks, as it aims to identify the most relevant and informative features from a given dataset. In this paper, the random forest technique was employed to generate a subset of the most important features from a dataset comprising nine initial features. The random forest technique, a widely used ensemble learning method, was utilized to rank the importance of features within the dataset. By constructing multiple decision trees using bootstrapped subsets of the original data, the algorithm effectively selects the most discriminative features based on their ability to improve prediction accuracy. Leveraging the robustness and ability to handle high-dimensional data inherent in the random forest algorithm, the features were ranked based on their importance scores. Through this process, a subset of the top-ranked features was successfully identified.

Number #	Attribute	Domain
1	Sample code number	Id number
2	Clump Thickness	1–10
3	Uniformity of Cell Size	1–10
4	Uniformity of Cell Shape	1–10
5	Marginal Adhesion	1–10
6	Single Epithelial Cell Size	1–10
7	Bare Nuclei	1–10
8	Bland Chromatin	1–10
9	Normal Nucleoli	1–10
10	Mitoses	1–10
11	Class	2 for Benign, 4 for Malignant

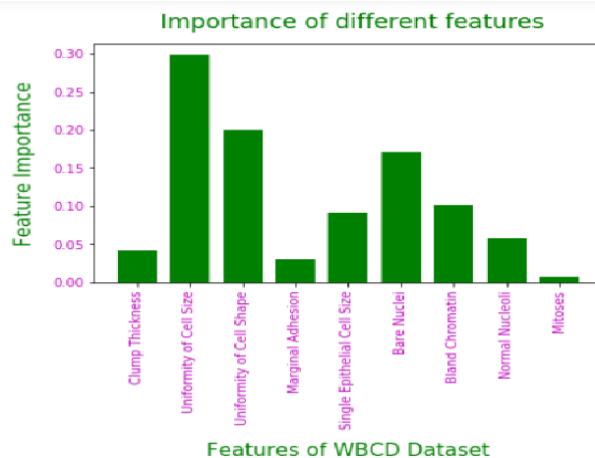**Table 1. Attributes of WBCD dataset**

Fig. 4. Representing the importance of features

Algorithm: Fusion Cascade: A Cascading Ensemble with Custom Subset Generation and Multi-Level Fusion

s1, s2, s3, s4, s5 – subsets generated from training set, VPL1 – validation predictions from level 1, VPL2 – validation predictions from level 2, TPL1 – test predictions from level 1, TPL2 – test predictions from level 2, TSL2 – training set for level 2, ESL2 – test set for level 2, RF – random forest, TF6 – top 6 features, CVP – combined validation predictions of level 1 and level 2, CTP – combined test predictions

of level 1 and level 2, FTrP – final train predictions, FTeP – final test predictions

Input: Dataset-[WBCD with 699 samples and 9 features, 458 are benign and 241 are malignant] Output: [confusion matrix, accuracy, sensitivity, specificity]

Step 1: Load Dataset (DT)

Step 2: Pre-processing of Dataset

Step 3: Dividing dataset into testing and training ratio(30:70)

Step 4: Generate 5 subsets (s1, s2, s3, s4, s5) from training set

Step 5: Training first level classifiers(DT, LR, KNN, SVM, GNB) on subsets(s1, s2, s3, s4, s5) with 10 fold cross validation

    5.1: VPL1 = combine (validations predictions from all models)

    5.2: TPL1 = combine (test predictions from all models)

Step 6: TF6 = RF (DT)

Step 7: TSL2 = combine (VPL!, TF6), ESL2 = combine (TPL1, TF6)

Step 8: Training second level classifiers(Adaboost, GBM, RF) on TSL2 and validated on ESL2

    8.1: VPL2 = combine (validations predictions from all models)

    8.2: TPL2 = combine (test predictions from all models)

Step 9: CVP = combine (VPL1, VPL2), CTP = combine (TPL1, TPL2)

    9.1: FTrP = DNN (CVP)

    9.2: FTeP = DNN (CTP)

Step 10: compute [confusion matrix, accuracy,specificity,sensitivity]

Step 11: evaluation and analysis

4. EXPERIMENTS AND RESULTS

This section presents the results, its analysis, the strategies applied in designing the model and the importance of the strategies. The WBCD dataset comprises 699 samples, including 458 benign cases and 241 malignant cases. For the CECSMF model, this dataset is applied to train and evaluate the model. In the first level, five different classifiers are employed, each trained on distinct subsets generated from the training set. Great care is taken in preparing these subsets to ensure fairness and representative sampling. Specifically, 50% of the training set samples are allocated to each of the subsets, while the remaining samples are randomly sampled. Importantly, the distribution of benign and malignant cases in the subsets mirrors that of the original training set. To maintain diversity and avoid overfitting, no sample is repeated more than three times within a subset. Furthermore, all samples in the training set are allocated to one of the subsets, and each subset is of equal size to the training set.

After training the first-level classifiers on subsets using a 10-fold cross-validation approach, their performance and predictive capabilities are evaluated on the test set. To ensure reliable and unbiased results, a meticulous approach is followed in subset preparation and evaluation. The validation predictions of all first-level classifiers are combined based on the Sample Code Number, and then merged with the top 6 features obtained from Random Forest analysis. Similarly, the test predictions of all classifiers are combined and concatenated with the top 6 features. Based on these first-level predictions, the classifiers are evaluated, and the results are presented in Table 2. Using the combined validation predictions and top 6 features, the second-level classifiers are trained using a 10-fold cross-validation approach. Subsequently, the test predictions, combined with the top 6 features, are utilized to

evaluate the performance of the second-level classifiers. The results of this evaluation are also presented in Table 2. This comprehensive approach, incorporating subset preparation, evaluation, and feature combination at each level, ensures reliable and unbiased outcomes for the CECSMF model. The presented Tables 1 and 2 provide insights into the performance of the classifiers at each level and demonstrate the effectiveness of our ensemble approach.

In the next step, the predictions from the first and second levels are combined to train and evaluate the third-level classifier. The predictions of the third-level classifier represent the final predictions of the CECSMF model. The results of the third-level classifier, including confusion matrices, ROC curves, accuracy, sensitivity, specificity, and AUC-ROC values, are presented in a comprehensive manner in this section. These results provide a holistic assessment of the performance and effectiveness of the ensemble model, allowing for a thorough evaluation of its predictive capabilities.

The proposed CECSMF model achieved promising results, as indicated by the accuracy, sensitivity, specificity, and AUC-ROC values obtained by each level of classifiers. At the first level, the classifiers demonstrated high accuracy values, ranging from 96.67% to 98.09%. This indicates that the individual classifiers were able to effectively classify the samples in the WBCD dataset. Similarly, the sensitivity values, which represent the ability to correctly identify malignant cases, ranged from 93.15% to 95.89%, showing the classifiers' capability to capture the presence of malignant instances.

Table 2. Results of the proposed model at different levels

Models/Metrics	First Level Learners					Second Level Learners			Third Level Learner
	Decision Tree	Logistic Regression	K-Nearest Neighbours	Support VectorMachine	Gaussian Naïve Bayes	Gradient Boosting Machine	Random Forest	AdaBoost	Multi-Layer Perceptron
Confusion Matrix	[[136 1] [5 68]]	[[135 2] [3 70]]	[[136 1] [3 70]]	[[136 1] [3 70]]	[[135 2] [5 68]]	[[135 2] [0 73]]	[[136 1] [1 72]]	[[134 3] [0 73]]	[[136 1] [0 73]]
Accuracy	97.14	97.62	98.09	98.09	96.67	99.04	99.04	98.57	99.52
Sensitivity	93.15	95.89	95.89	95.89	93.15	100	98.63	100	100
Specificity	99.27	98.54	99.27	99.27	98.54	98.54	99.27	97.81	99.27
AUC-ROC	0.96	0.97	0.98	0.98	0.96	0.99	0.99	0.99	1

Additionally, the specificity values, measuring the ability to correctly identify benign cases, were generally high, ranging from 98.54% to 99.27%. Moving to the second level, the classifiers continued to exhibit exceptional performance. The accuracy values remained consistently high, ranging from 98.57% to 99.04%. Moreover, the sensitivity values were impressively high, with a minimum value of 98.63%, indicating the classifiers' ability to accurately identify malignant cases. The specificity values remained strong as well, ranging from 97.81% to 99.27%.

Finally, the third-level learner, represented by the Multi-Layer Perceptron (MLP) classifier, demonstrated exceptional performance across all metrics. The accuracy reached 99.52%, indicating the model's high overall prediction accuracy. Furthermore, both sensitivity and specificity values were perfect, indicating the MLP's ability to correctly identify both benign and malignant cases. The AUC-ROC value of 1 indicates that the MLP classifier achieved a perfect balance between sensitivity and specificity.
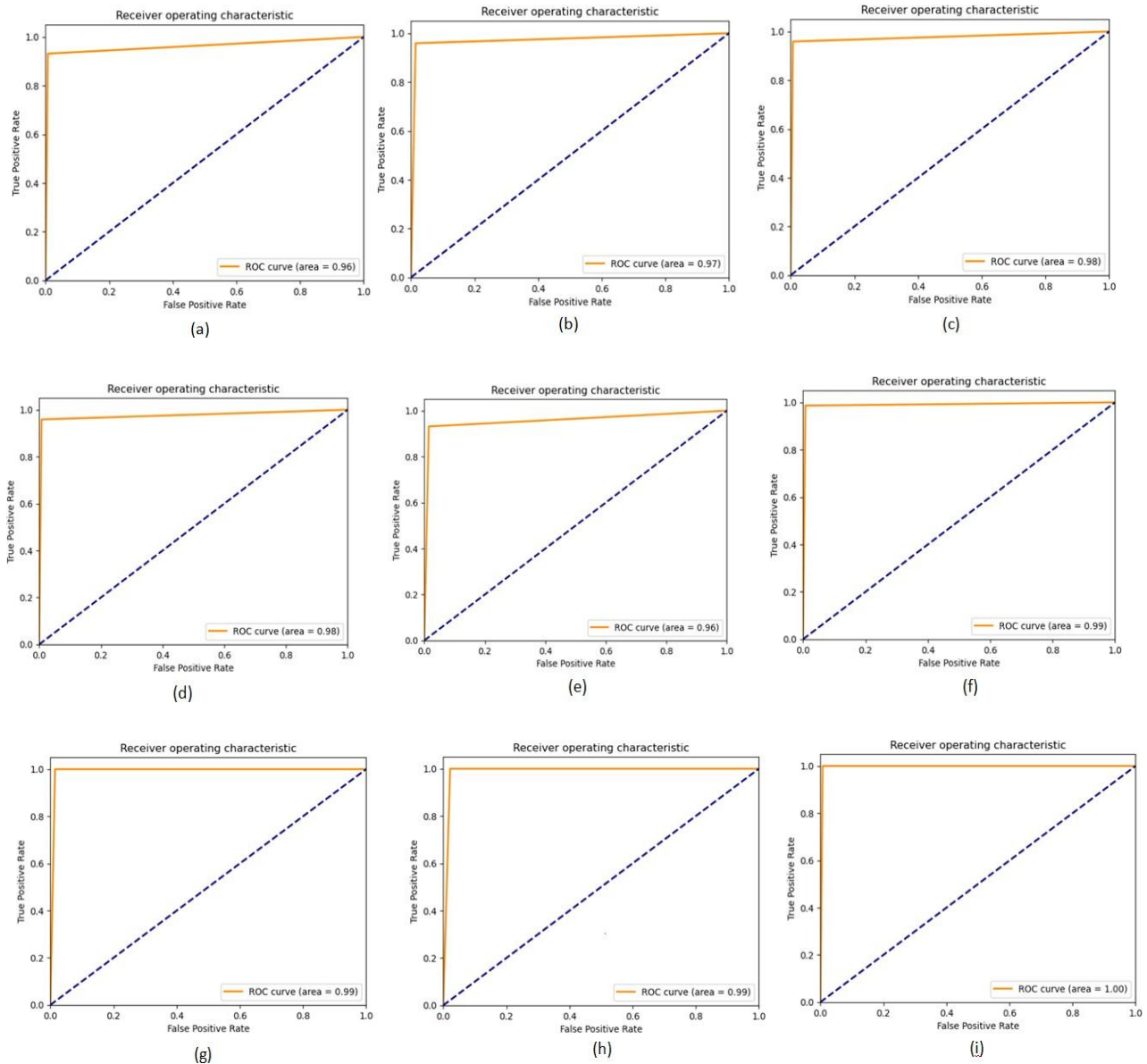
Fig. 5 ROC curves a)DT b)LR c)KNN d)SVM e)GNB f)RF g)GBM h)AdaBoost i)MLP

Overall, the results demonstrate the effectiveness of the proposed CECSMF model. The ensemble of diverse classifiers at each level contributed to the model's high accuracy, sensitivity, specificity, and AUC-ROC values. This suggests that the model is capable of accurately distinguishing between benign and malignant cases in the WBCD dataset, making it a promising approach for breast cancer diagnosis.

Distinct classifiers at different levels

Different classifiers have distinct strengths and weaknesses in capturing patterns and relationships within the data. By incorporating a diverse set of classifiers, each level can leverage the unique abilities of different models to enhance the overall predictive performance. This strategy allows for error correction and reduction. If one classifier at a particular level makes an incorrect prediction, the other classifiers can compensate for that error in subsequent levels. This helps to mitigate the impact of individual classifier weaknesses and enhances the ensemble's overall accuracy and robustness.

This strategy can also help to combat overfitting, a common issue in machine learning. Each classifier's predictions are based on its own learned representation of the data, reducing the risk of over-reliance on a single model's biases or idiosyncrasies. The ensemble can average out individual model biases, leading to improved generalization performance and reduced overfitting. The use of different classifiers adds stability to the ensemble. If a single classifier is highly sensitive to small changes in the training data or initialization, the impact on the ensemble's predictions is reduced when multiple classifiers with different sensitivities are combined. This can result in a more reliable and robust ensemble that produces consistent predictions across different variations of the data.

Training classifiers on different subsets

When training classifiers on different subsets, each subset represents a distinct portion of the data with potentially unique patterns and characteristics. By exposing classifiers to diverse subsets, they can learn different aspects of the data, leading to a more comprehensive understanding of the underlying relationships. This can enhance the ensemble's ability to generalize well to unseen data by capturing a broader range of patterns and reducing the risk of overfitting to specific instances or biases present in a single dataset.

By combining classifiers trained on diverse subsets, the ensemble can leverage a wider range of perspectives and insights, leading to improved robustness and performance. Training classifiers on different subsets allows for ensemble calibration, enables the ensemble to detect and correct errors more effectively. If a particular subset contains noisy or misleading data, the classifiers trained on other subsets can identify and mitigate these errors during the aggregation process. This error detection and correction mechanism enhances the ensemble's robustness and reduces the risk of individual classifiers making significant prediction errors.

Combination of important features with predictions

Incorporating top features along with previous level predictions provides a more comprehensive representation of the data. While previous level predictions capture the collective insights of the ensemble's classifiers, including top features allows the subsequent classifiers to leverage the most informative and relevant aspects of the data. This combination can enhance the discriminative power of the features and improve the overall representation of the input data.

Combining previous level predictions with top features enables synergistic information fusion. The predictions from the previous level capture the ensemble's collective knowledge, while the top features provide additional discriminative information. By combining these two sources of information, the subsequent classifiers can leverage the complementary strengths of both the predictions and the features, leading to a more accurate and robust model.

Combination of two level predictions

Training the third level classifier on both the first and second level predictions in the stacking ensemble brings together multiple levels of information, enables capturing complex patterns, improves generalization, enhances predictive power, increases robustness against errors and biases, and facilitates the integration of informative features. These benefits contribute to a more powerful and reliable ensemble model, capable of leveraging the collective knowledge of the ensemble and producing accurate

predictions on unseen data.

Table 3. presents comparison between previous and proposed research

S.No	Author(Yr)	Method	Accuracy %
1	In 2012 Hsieh SL	Single NF	94.28
		Single KNN	96.42
		Single QC	94.5
		NFE(Ensemble)	94.56
		KNNE(Ensemble)	96.42
		QCE(Ensemble)	96.57
		NF KNN QC(Ensemble	97.14
2	In 2015 Onan A	BayesNet(Dagging)	97.66
		Single RIPPER	95.17
3	In 2015 Bashir S	NB+DT(gini)+DT(IG)+SVM	97.42
4	In 2017 Yavuz E	Neural Networks ensemble	96.43
5	In 2018 Wang H	WAUCE(Ensemble)	97.68
6	In 2019 Kwon H	GBM(Ensemble)	97.96
7	In 2020 Dhanya R	Boosting with feature selection	97.37
8	In 2020 Abdar M	SV-Naives Bayes-3-meta classsifier	98.07
		SV-BayesNet-3-Meta classifier	
9	In 2021 Chaurasia V	LR(Ensemble)	98.96
10	Present Research	Three level stacking ensemble	99.52

CECSMF ensemble model

Three levels for the ensemble strikes a harmonious balance among several critical considerations, encompassing model complexity, computational efficiency, ensemble diversity, and interpretability. This decision allows the ensemble to effectively amalgamate insights across multiple levels while maintaining an optimal degree of complexity and mitigating potential drawbacks inherent in using fewer or more levels. Having three levels facilitates the integration of diverse models and their corresponding predictions. Each level offers a distinct set of classifiers, ensuring ensemble diversity by capturing varied perspectives and exploiting different aspects of the data. This diversity enhances the ensemble's collective wisdom and reduces the risk of bias, resulting in more robust and reliable predictions.

As the number of levels increases, the computational complexity and training time of the ensemble escalate. However, opting for three levels strikes an optimal compromise, balancing the need for ensemble performance and computational efficiency. This allows for efficient utilization of computational resources while maintaining an acceptable level of model complexity.

The stacking ensemble strives to strike a balance between reducing bias and managing variance in predictions. With three levels, the ensemble leverages the collective knowledge from multiple levels to mitigate bias while avoiding excessive model complexity that could lead to overfitting and increased variance. This balance enhances the ensemble's generalization capability, enabling it to make accurate predictions on unseen data.

Strategies in subset generation

First 50% of training set samples are allocated to all subsets ensure that each subset initially contains a representative portion of the training data. This step helps establish a foundation of consistent samples across all subsets. From the remaining 50% of the training set, samples are randomly selected and added to all subsets until the size of each subset matches the size of the training set. This random selection helps maintain diversity within each subset and ensures that all samples have an equal chance of being included.

By maintaining the same distribution of benign and malignant cases in the subsets as in the training set, it is ensured that each subset is representative of the overall class distribution. This is important for training models that can generalize well to unseen data. Limiting the number of times a sample can repeat in a subset (up to 3 times) helps prevent over-representation of certain data points. This ensures that the subsets are balanced and that no individual sample has an overly dominant influence on the training process. Every subset is of the same size as the training set, this helps maintain consistency in the training process. It ensures that each subset has a comparable number of samples, allowing fair evaluation and comparison of models trained on different subsets.

Overall Analysis

Using the above specified strategies, a three-level stacking ensemble model was thoughtfully designed, thoroughly trained, and meticulously evaluated. The results of the proposed model are presented in Table 2 and Table 3. This ensemble model leverages diverse classifiers, controlled random sampling, cross-validation, feature selection, multilevel fusion and a neural network to create a powerful and robust prediction system. The use of multiple levels and distinct classifiers enables the ensemble to capture different perspectives, exploit collective insights, and effectively learn complex patterns in the data, leading to improved predictive performance.

5. CONCLUSION AND FUTURE WORK

In conclusion, the proposed ensemble model demonstrated significant improvements in predictive power and robustness compared to individual classifiers. By leveraging a diverse set of classifiers in the first level and incorporating advanced machine learning techniques in subsequent levels, the ensemble model achieved enhanced performance. The amalgamation of different perspectives, diverse classifiers, and the utilization of top-ranked features contributed to the ensemble's success.

Future work can focus on exploring additional ensemble diversity, conducting comprehensive feature engineering; optimizing model hyper parameters and evaluating the ensemble in real-world

applications to further enhance its capabilities and applicability. Overall, the stacking ensemble model offers a promising approach for improving classification tasks.

REFERENCES

1. Hsieh SL, Hsieh SH, Cheng PH, Chen CH, Hsu KP, Lee IS, Wang Z, Lai F. Design ensemble machine learning model for breast cancer diagnosis. Journal of medical systems. 2012 Oct;36(5):2841-7.

2. Onan A. On the performance of ensemble learning for automated diagnosis of breast cancer. In Artificial Intelligence Perspectives and Applications 2015 (pp. 119-129). Springer, Cham.

3. Bashir S, Qamar U, Khan FH. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Quality & Quantity. 2015 Sep;49(5):2061-76.

4. Yavuz E, Eyupoglu C, Sanver U, Yazici R. An ensemble of neural networks for breast cancer diagnosis. In2017 International Conference on Computer Science and Engineering (UBMK) 2017 Oct 5 (pp. 538-543). IEEE.

5. Kwon H, Park J, Lee Y. Stacking ensemble technique for classifying breast cancer. Healthcare informatics research. 2019 Oct;25(4):283.

6. Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm forbreast cancer diagnosis. European Journal of Operational Research. 2018 Jun 1;267(2):687-99.

7. Dhanya R, Paul IR, Akula SS, Sivakumar M, Nair JJ. F-test feature selection in Stacking ensemble model for breast cancer prediction. Procedia Computer Science. 2020 Jan 1;171:1561-70.

8. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters. 2020 Apr 1;132:123-31.

9. Chaurasia V, Pal S. Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer. SN Computer Science. 2021 Apr;2(2):1-3.

10. Jakhar AK, Gupta A, Singh M. SELF: a stacked-based ensemble learning framework for breast cancer classification. Evolutionary Intelligence. 2023 Jan 31:1-6.

11. Kadam VJ, Jadhav SM, Vijayakumar K. Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. Journal of medical systems. 2019 Aug;43(8):263.

12. Cai T, He H, Zhang W. Breast cancer diagnosis using imbalanced learning and ensemble method. Applied and Computational Mathematics. 2018;7(3):146-54