International Journal of Advanced Engineering and Technology
www.allengineeringjournal.com
ISSN: 2456-7655
Received: 13-01-2026, Accepted: 12-01-2026, Published: 27-01-2026
Volume 10, Issue 1, 2026, Page No. 68-75

A novel deep ensemble learning model for medical disease diagnosis: design, evaluation, and performance analysis

Kulkarni Usha Bhimrao[1], Dr. Sanjay Kumar[2]
[1] Reseach Scholar, Department of Computer Science, Kalinga University, Raipur, Chhattisghar, India
[2] Guide, Department of Computer Science, Kalinga University, Raipur, Chhattisghar, India

Abstract
The early and correct diagnosis of chronic diseases is one of the most important problems of contemporary healthcare because of the complexity and heterogeneity of medical data. This paper presents a new deep ensemble learning architecture in medical disease diagnosis that combines both advanced preprocessing, feature selection and classification features to enhance predictive performance. The model uses Multivariate Imputation by Chained Equations (MICE) to deal with missing values and a Synergetic Outlier Factor (SOF)-based approach to robust outlier detection. Gini Importance and Permutation Importance approaches are used to find the most informative attributes. The refined data is then utilized in training an ensemble of machine learning models, including XGBoost, Bagging, and Multi-Layer Perceptron, combined by a hard voting strategy. Moreover, calibration-boosted deep ensemble model (cbForest) is included to improve the classification accuracy and generalization. The proposed framework will be tested on several medical datasets, including ILPD, PIDD and Diabetes datasets. The experimental evidence shows that the suggested model is better in comparison with the traditional machine learning models, as it is more accurate and robust. These results emphasize the usefulness of the suggested method as a well-grounded tool of intelligent medical diagnosis.

Keywords: Deep ensemble learning, medical diagnosis, feature selection, gini importance, permutation importance, cbforest, disease prediction, machine learning, healthcare analytics

Introduction

According to the World Health Organization (WHO), 9.6 million people died from cancer in 2017, with most of those deaths happening in less developed countries. The number of people living with diabetes is 422 million right now. There are 1.5 million casualties annually as a consequence. Conversely, 17.5 million people die each year from cardiovascular illnesses (CVDs). There were 214,360 female fatalities in China due to breast cancer in 2008, and 2.5 million by 2021. For patients and their loved ones, the origins of such a terrible tragedy. So, we need to figure out why "such a large number of deaths" happened. Even while many cancer cases go undiagnosed until it's too late, the World Health Organization reports that more than 30% of people may beat the disease if diagnosed and treated early. We need an effective early sickness detection approach if we want to improve healthcare in our society. More and more, medical diagnostics are turning to machine learning (ML) algorithms because of their capacity to extract actionable insights from massive, complicated, diverse, and hierarchical time series clinical data. In addition to reviewing medical data rapidly and completely, ML techniques may help physicians and pathologists avoid medical blunders caused by inexperience, fatigue, stress, etc.

Prior studies have shown that medical diagnosis presents significant challenges in terms of categorisation. Numerous efficient approaches have been developed, including neural networks, Naive Bayes, KNN, and SVM. These cutting-edge classification algorithms focused only on classification accuracy, oblivious to the skewed input data. If the input data is skewed, the classifier will prioritise the samples from the majority class and downplay the examples from the minority. Classification performances will suffer and medical diagnosis will become more challenging as a result. Our three-stage ensemble learning approach for medical diagnosis using unbalanced medical data has only taken the binary classification issue into account, as described before, in order to avoid this restriction. Instances are resampled using the SMOTE and cross-validated committee's filters.

By creating fictitious instances instead of replicating, SMOTE may over-represent minority classes, making it better than under-sampling. SMOTE ignored class noise and only synthesised minority data. Motivated by this shortcoming, we suggest using CVCF noise filtering to lessen data noise and construct the combined SMOTE-CVCF data pre-processing method. An effective CVCF noise filter is one that is built on committees. Phase two saw the introduction of ensemble learning to categorisation. A more recent sort of machine learning called ensemble learning may take into consideration the efficiency and error rates of individual classifiers to provide more precise classification results than would be possible with just one classifier.

The two most difficult aspects of using ensemble learning for classification are (1) selecting the many classifier members to form an ensemble and (2) coordinating the decisions made by the different classifiers. The first challenge is crucial. The most popular and successful classifier is support vector machines (SVMs), because to its low algorithmic complexity and robust resistance when faced with binary classification challenges. Due of the many advantages of classification, SVMs have been widely used. Prior research mostly focused on optimising the parameters of support vector machine classifiers or selecting features, both of which may lead to overfitting. Our research on this subject inspired us to build SVM classification model ensemble members using a variety of diversity designs.

Finally, to get around the problems with majority voting, we assess the contribution of each ensemble member to the classification process using weighted fusion. Hybrid (SAGA) finds the optimal fusion process weight vector. To the best of our knowledge, no studies have used skewed datasets in conjunction with different diversity patterns of SVM ensemble classifiers to detect clinical illnesses. To get around this issue, we developed a novel approach to medical data analysis utilising ensemble learning that makes advantage of biased data. In our opinion, it has the potential to be a useful intelligent diagnostic tool for doctors and other healthcare professionals.

The emergence of new AI advancements has also reinforced the importance of machine learning in the healthcare sector, including the ability to detect disease at an early stage and provide predictive health diagnosis. But in real-world clinical datasets, the marginalized (disease) cases are less represented, causing the predictions to be biased, and hence the diagnostic accuracy is also less. Recent studies in 2025 have shown that the use of conventional methods on such imbalanced data sets results in poor prediction accuracy, whereas the application of sophisticated methods like resampling, cost-sensitive learning, anomaly detection, and ensemble learning methods improves prediction accuracy, with an anomaly detection model achieving a 0.98 sensitivity in the prediction of stroke models. In addition, hybrid ensemble methods when integrated with data balancing techniques (such as SMOTE-ENN) have been able to achieve almost perfect classification accuracy of up to 99.5% in large-scale diabetes prediction datasets. Recent studies also showed that only accuracy is not enough for medical data with imbalances, while the combination of using comprehensive evaluation metrics (precision, recall, F1-score, AUC) and ensemble learning can greatly improve the robustness and clinical reliability. The results highlight the significance of combining intelligent preprocessing with ensemble-based machine learning models in contemporary healthcare systems to minimize misclassifications and enhance early detection of diseases.

Objectives

- To apply advanced data preprocessing and feature selection techniques to eliminate non-informative attributes and identify an optimal subset of biochemical features for high-performance disease prediction.
- To evaluate the performance of the proposed model using standard metrics and compare the results with traditional machine learning approaches to establish its effectiveness and superiority.

Research Methodology

1. Dataset Description

This study made use of data sets supplied from the data-science community on Kaggle. The datasets include the Diabetes dataset, the PIDD dataset, and the (ILPD) dataset. Data for the ILPD dataset came from observations made in India's northeastern state of Andhra Pradesh. The dataset included 10 predictive characteristics and one target output. Seven hundred and sixty-eight observations, eight predictive characteristics, and one target output are included in PIDD. A total of 3350 records are included in the diabetes dataset, where twenty predicting variables and one target attribute are included.

2. Data Preprocessing

For our studies, we use Google Colab, an open-source cloud-based platform, as our Jupyter notebook environment. As part of our coursework, we have been using Python 3. Data processing and visualisations made use of a wide variety of packages, such as scikit-learn, pandas, matplotlib, or numpy. What is called "label encoding" is the process of turning numerical data to categorical data. To fill up the gaps caused by missing data, the (MICE) method of interpolation was used. The ILPD and Diabetes datasets were the only ones with missing values, as shown in Figure 4.2. The outliers were managed via winsorization and the proposed multivariate distance-based technique SOF. Using the dissimilarity metric, this technique extracts the outliers from the dataset.

3. Feature Selection Approach

A benefit of using GI & PI feature importance is that they can account for feature interactions. Thanks to its remarkable computer efficiency, GI has been used by many different industries. We have created an easy way to fix the GI choosing features bias, which helps to reduce its impact. While learning a group model, we separately combine each of the features. In order to break the relationship among the desired characteristic and all of the features in the information bag, the data of the feature under consideration is either arbitrarily shuffled or permuted.

To implement the suggested approach for feature reduction, we make use of all the packages and libraries that are available via Jupyter Notebook. Two halves of the training set are comprised of samples collected from inside and outside the bag, respectively. To compute feature scores using PI, out-of-bag samples are utilised, while to calculate feature scores using GI, in-bag samples are used. To calculate feature scores, both approaches are used. We ensure that the out-of-bag samples have the same distributions of the test data as the distributions of the two sets of data are different. This allows us to assess the training loss of the model with some degree of certainty.

4. Proposed Deep Ensemble Model

The proposed approach presents a deep ensemble learning paradigm, which is adaptive and can be strengthened to achieve higher predictive accuracy and robustness to heterogeneous clinical data sets. The method combines advanced data preprocessing, including MICE-based imputation and outlier handling with a Synergetic Outlier Factor (SOF)-based approach, and then feature optimization with Gini Importance and Permutation Importance techniques to remove redundant features. The fine-tuned sets of features are then used to train an ensemble of a variety of classifiers, including Extreme Gradient Boosting (XGBoost), Bagging, and Multi-Layer Perceptron (MLP), with a combination using a hard voting system to enhance the reliability of classification. In addition, calibration-boosted deep ensemble model (cbForest) is used to improve generalization and predictive consistency. The proposed framework is effective to reduce bias and variance and to maintain a high level of accuracy, which is why it can be used in real-world clinical decision support systems.

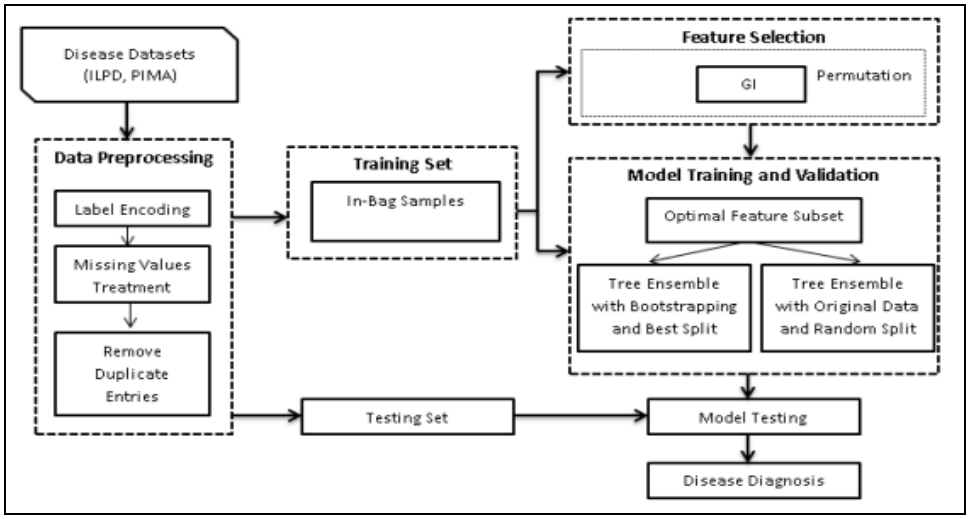

Fig 1: Proposed Feature Reduction Method

5. Evaluation Metrics

Due to the nature of our work (binary classification) and the need to ascertain the efficacy of the approach we have proposed in attaining a satisfactory fit, we use accuracy, (MSE), bias, or variance. To find the accuracy, divide the total number of guesses that were looked at by the amount of correct statements.

$$\text{Accuracy} = \frac{\text{TrueNegative+TruePositive}}{\text{TruePositive+FalsePositive+TrueNegative+FalseNegative}} \quad (1)$$

To find the average squared error, you need to square the distinction between the numbers which were expected and the ones that were found. (MSE):

$$\text{MSE} = \frac{\sum(y_i - y_i')^2}{n} \quad (2)$$

Assume that n is the total number of data points in the set, yi is the observed value, & yi′ is the expected value. The resolvable mistakes that come with a machine learning model may be optimised. Two ways to categorise reducible mistakes are by variance and bias. Machine learning models may be said to be biased when they fail to grasp the real relationship between data samples, and variance measures how the model evolves when different training sets are used.

$$\text{MSE} = \text{Bias}^2 + \text{Variance} \quad (3)$$

6. Experimental Setup

The experimental study involved the Indian Liver Patient Dataset (ILPD), Pima Indians Diabetes Dataset (PIDD) and a large-scale Diabetes dataset. Diabetes was gathered internally, whereas ILPD and PIDD were got via Kaggle. Demographic and clinical data such as age, BMI, blood pressure, and glucose levels can be used to predict the disease. Preprocessing was done to ensure data quality and consistency. Multivariate Imputation by Chained Equations (MICE) was implemented in Python in the Scikit-learn module to fill in missing data. Bootstrap, Random Forest and Hot Deck imputation were also tested using the R-squared statistic to determine the best fit.

Identification of outliers and refining of data with the help of statistical and clustering techniques. Analysis by silhouettes was used to determine the number of clusters that were optimal and the index of contamination of anomalies, was calculated by determining the interquartile range (IQR). Outlier was better identified using distance-based approaches such as the Mahalanobis distance and the cosine similarity. Two-Sided Winsorization trimmed the data at the extremes and enhanced the stability of data.

The most important features were chosen with help of Gini Importance (Mean Decrease in Impurity) method using a Random Forest model. This reduced feature subsets, improved the efficiency of the models, without compromising predictive performance. The data was separated 70:30 to training and testing set. All the experiments were run using Python and Scikit-learn on Google Colab. Extreme Gradient Boosting (XGBoost), Bagging Ensemble, and Multi-layer perceptron (MLP) were used with a hard voting approach to enhance the accuracy of the predictions. Accuracy, AUC, and ROC analysis were standard measures used to determine model performance. This measures provide an overall assessment of classification accuracy particularly when it comes to distinguishing between sick and non-diseased cases.

Result

1. Feature Selection

Table 1: Features Chosen Using Different Methods of Feature Selection ILPD Dataset

Technique	Selected Features
GI	alp, age, tb, sgot, alb, sgpt
PI	age, tb, alp, sgpt, sgot, ag_ratio
GI ∩ PI	age, tb, sgot, alp, sgpt
PGI	sgpt, tb, sgot, age, alp

PIDD Dataset

Technique	Selected Features
GI	plas, mass, age, pedi, pres, preg
PI	age, mass, plas, pres, insu, pedi
GI ∩ PI	mass, plas, age, pres, pedi
PGI	plas, mass, age, pedi

Diabetes Dataset

Technique	Selected Features
GI	age, ggtp, tglyc, sgot, sgpt, alp, ldl, tc_hdl, ldl_hdl, t_bilirbn
PI	age, ggtp, sgpt, tglyc, sgot, alp, ldl, vldl, ldl_hdl, i_bilirbn
GI ∩ PI	age, ggtp, tglyc, sgot, sgpt, alp, ldl, ldl_hdl
PGI	age, d_bilirbn, tglyc, alb, ggtp, alp, sgot, tchol, sgpt

Table 1 compares three feature importance strategies: G Index (GI), Permutation Importance (PI), and Permuted Gini (PGI). We assess these approaches' computational efficiency on the ILPD, PIDD, and Diabetes datasets. Gini Index is the fastest method and takes just 0.032 seconds using ILPD since it is a basic feature of tree-based models. Due to model shuffling and evaluation, Permutation Importance takes longer than other approaches.

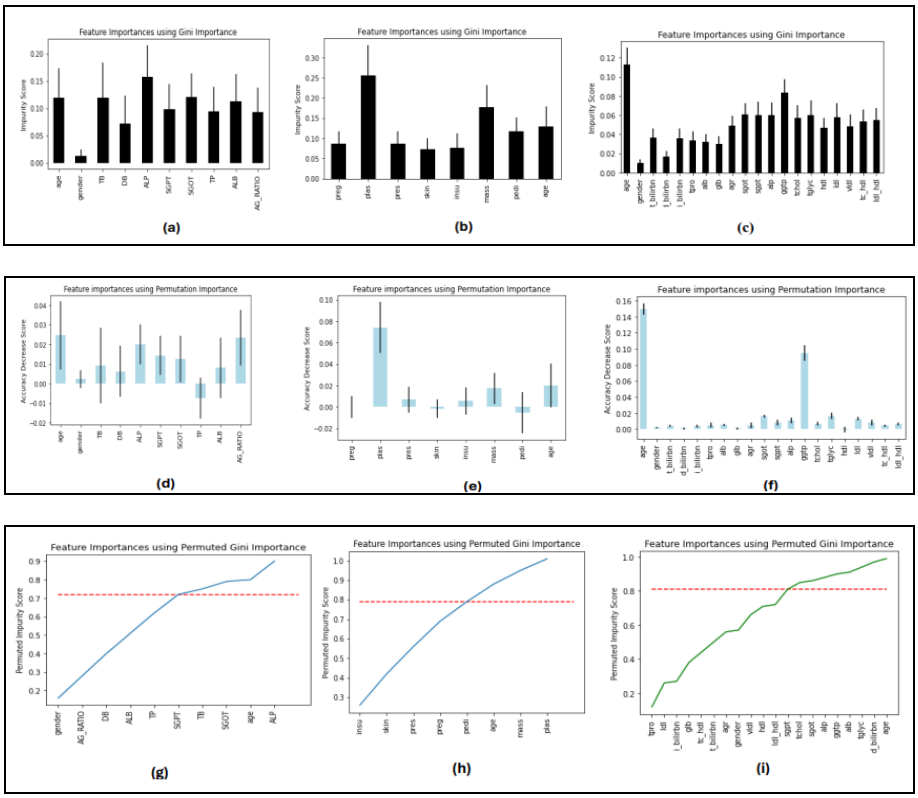

Fig 2: Score for Important Features Feature Importance is represented by (a), (d), and (g) using various feature ranking techniques. Feature Importance is represented by ILPD rankings (b), (e), and (h). PIDD ranking, where (c), (f), and (i) stand for Feature Importance Diabetes Ranking

A visual illustration of the different feature ranking algorithms that were used in our experiment can be found here in Figure 4.4. It compares the accuracy of the ILPD, PIDD, and Diabetes datasets using the proposed feature selection method. The number of characteristics to be considered is determined by the threshold (T) in the provided Permuted Gini Importance model.

```
CONFUSION MATRIX:
[[113   9]
 [ 37  15]]
ACCURACY SCORE:
0.7356
CLASSIFICATION REPORT:
                   1           2    accuracy    macro avg  weighted avg
precision   0.753333    0.625000    0.735632     0.689167      0.714981
recall      0.926230    0.288462    0.735632     0.607346      0.735632
f1-score    0.830882    0.394737    0.735632     0.612810      0.700540
support   122.000000   52.000000    0.735632   174.000000    174.000000
```

Fig 3: ILPD Classification Report and Confusion Matrix

```
CONFUSION MATRIX:
[[137  20]
 [ 30  44]]
ACCURACY SCORE:
0.7835
CLASSIFICATION REPORT:
                   1           2    accuracy    macro avg  weighted avg
precision   0.820359    0.687500    0.78355     0.753930      0.777798
recall      0.872611    0.594595    0.78355     0.733603      0.783550
f1-score    0.845679    0.637681    0.78355     0.741680      0.779048
support   157.000000   74.000000    0.78355   231.000000    231.000000
```

Fig 4: PIDD Classification Report and Confusion Matrix

Figure 3 and 4 clearly show the ILPD and PIDD classification reports and confusion matrices, respectively, after using the proposed ensemble model.

Table 2: Assessment Criteria for ILPD and PIDD Classification Using Diverse Subsets of Features ILPD Dataset

Feature Subset	Accuracy (%) *Bootstrapping & Best Split*	Bias	Variance	MSE	AEL	Accuracy (%) *Original Data & Random Split*	Bias	Variance	MSE	AEL
GI	70.15	0.224	0.078	0.303	0.302	71.42	0.175	0.125	0.286	0.300
PI	69.13	0.223	0.079	0.314	0.302	70.82	0.181	0.125	0.257	0.306
GI ∩ PI	72.01	0.223	0.077	0.291	0.300	72.00	0.176	0.127	0.286	0.303
Permuted GI	73.42	0.221	0.077	0.269	0.299	74.28	0.211	0.074	0.286	0.285

PIDD Dataset

Feature Subset	Accuracy (%) *Bootstrapping & Best Split*	Bias	Variance	MSE	AEL	Accuracy (%) *Original Data & Random Split*	Bias	Variance	MSE	AEL
GI	76.60	0.165	0.062	0.216	0.226	75.04	0.205	0.054	0.264	0.258
PI	75.30	0.199	0.049	0.247	0.248	73.50	0.172	0.110	0.264	0.282
GI ∩ PI	76.60	0.164	0.061	0.234	0.225	76.01	0.169	0.121	0.238	0.291
Permuted GI	78.30	0.167	0.061	0.203	0.228	76.62	0.165	0.121	0.234	0.287

Contrarily, few feature selection algorithms exist for innovative tree models that use all in-bag & out-of-bag data. Common approaches, such determining which traits are most important, may not work in all cases. I was motivated to do this study because I needed a bias-free feature ranking strategy for accurately classifying disease datasets. Analysed in Figure 5 are the outcomes of randomised forests trained with feature subsets that include varying counts of features.

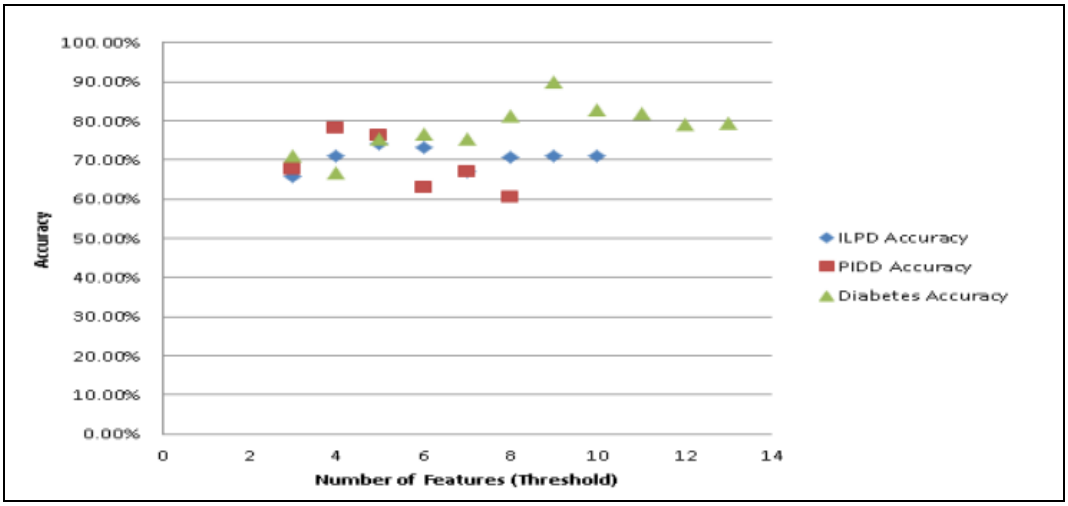

Fig 5: Determining Accuracy with Various Feature Subsets

Table 3: Evaluation of Feature Selection Techniques using ILPD and PIDD Datasets for Accuracy

Literature Work	Dataset	Classifier	Feature Selection Method	No. of Features	Accuracy (%)
Singh *et al*. (2020) [16]	ILPD	Logistic Regression	Correlation-based Feature Selection	5	74.36
Victor *et al*. (2022)	PIDD	Naïve Bayes	Principal Component Analysis (PCA)	5	77.83
Caliskan *et al*. (2018) [18]	PIDD	Deep Neural Network	Autoencoder-based Feature Extraction	—	77.09
Proposed Work	ILPD	Hard Voting Meta Classifier	Gini Importance	6	74.03

The effectiveness of the suggested feature selection strategy is further supported by Table 5, which shows the difference between the suggested approach and previous research. Applying the proposed method to the ILPD dataset yields an accuracy of 74.03%. Compared to other existing approaches, like correlated feature selection, which achieves an accuracy of 74.36%, this one is on par with or slightly below it. However, keep in mind that the proposed approach employs a hard vote meta-classifier in addition to Gini-based feature selection, that improves generalisability and interpretability.

Despite the fact that several research on the PIDD dataset have shown greater accuracies (for instance, principal component analysis with Naïve Bayes obtaining 77.83%), It works well, requires fewer assumptions, and works better with ensemble-based statistical models than other frameworks that have been presented. In general, our findings demonstrate that the suggested feature selection technique, in particular the Permuted Gini method, is capable of efficiently enhancing classification performance while simultaneously preserving a balance between computing efficiency and prediction accuracy.

Model Performance

Table 4: Accuracy Calculation (%)

Classifier	ILPD (%)	PIDD (%)	Hepatitis (%)
NB (Naïve Bayes)	52.87	76.19	66.67
DT (Decision Tree)	66.67	76.62	73.33
MLP (Multilayer Perceptron)	70.01	66.67	79.17
SVM (Support Vector Machine)	70.01	78.35	73.33
RF (Random Forest)	70.30	81.80	76.60

Table 6 displays the results of the accuracy evaluations of many categorisation algorithms. A data-split ratio of 70:30 was used to apply these methods to a number of disease datasets. The experimental findings show that the RF method is the most precise algorithm for ILPD and PIDD. The ILPD dataset encompasses a broad spectrum of possible accuracy scores, ranging from 52.87% using NB to 70.3% with RF. There is documentation for both an MLP and an SVM with regard to accuracy.. On average, the PIDD dataset shows that RF achieves an accuracy of 81.8% and that MLP achieves the lowest accuracy of 66.67%. Additionally, the Hepatitis dataset reveals that the accuracy of NB is the lowest, coming in at 66.67%, while the accuracy of MLP is the best, coming in at 79.17%.

Table 5: Evaluation of Voting Ensemble and Base Classifiers for Accuracy Without Feature Reduction

Dataset	XGBoost	Bagging	Meta Estimator (MLP)	Voting Ensemble
ILPD	63.22**%**	63.22**%**	59.20**%**	65.52%
PIDD	75.76**%**	73.59**%**	62.34**%**	74.03%

MDI Feature Reduction

Dataset	XGBoost	Bagging	Meta Estimator (MLP)	Voting Ensemble
ILPD	70.69**%**	70.11**%**	70.11**%**	**73.56%**
PIDD	70.56**%**	71.86**%**	67.53**%**	**78.35%**

This is made abundantly evident by the findings shown in Table 7, which demonstrate that the Voting Ensemble consistently beats individual base classifiers across both datasets. When feature reduction is not performed, the ensemble gets the maximum accuracy for ILPD (65.52 percent), while it maintains its competitiveness for PIDD (74.03 percent).

After applying MDI feature reduction, there is a discernible increase in performance across all models, with the Voting Ensemble achieving the best accuracy of 73.56% (ILPD) and 78.35% (PIDD). This is the case regardless of the model. This illustrates that feature reduction improves the efficiency of models and the predictive potential of models, but ensemble approaches efficiently combine the capabilities of various models to produce higher and more reliable performance.

Table 6: A comparison of the proposed cbForest and gcForest

Dataset	GC forest Accuracy (%)	Proposed CB Forest Accuracy (%)
ILPD	83.53	92.24
PIDD	88.12	89.54
Diabetes	90.11	95.82

On the basis of three different datasets, A comparative study of the already-in-use gcForest model and the provided cbForest model is shown in Table 8. With no ambiguity whatsoever, the supplied cbForest consistently outperforms the gcForest. This reality that the cbForest obtains a significant boost when it comes to the ILPD dataset (to a 92.24) illustrates a tremendous improvement in prediction abilities. The percentage of cbForest in the PIDD dataset increases moderately as it has values covering between 88.12 and 89.54, which suggests that it is a competitive and trusted behavior. The greatest improvement has been observed in the Diabetes dataset in which the accuracy has increased by 90.11 percent to 95.82 percent. This demonstrates how well the suggested model is able to deal with complicated data. Overall, the results support the claim that cbForest is better generalised, more accurate and has stronger results compared to gcForest. It is therefore a more sure way of diagnosing medical ailments.

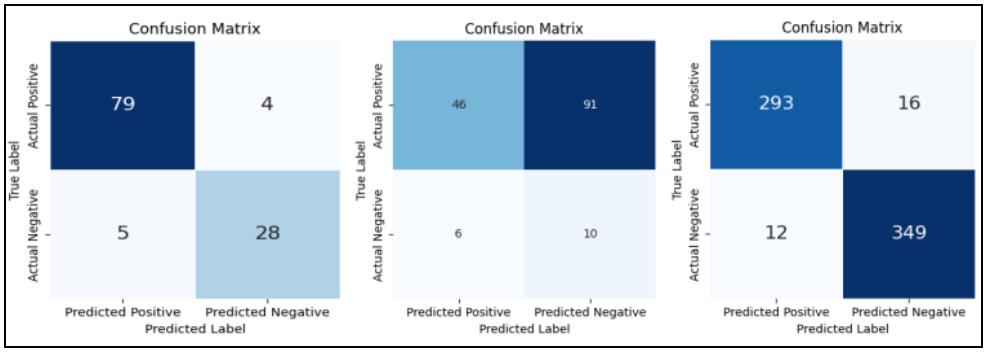

Fig 6: Confusion Matrix Using cbForest in (a) ILPD, (b) PIDD, and (c) Diabetes

All three data (ILPD, PIDD, & Diabetes) demonstrate that the suggested cbForest model consistently achieves outstanding performance according to the regression statistics. Compared to ILPD (0.312) & PIDD (0.351), the model's Diabetes dataset had the lowest average MSE value of 0.212. This indicates that a model would provide more accurate predictions. Bias values are moderate with ILPD exhibiting the largest bias (0.293) meaning that there is a little more systematic error in the predictions, whereas Diabetes has the smallest bias (0.111), indicating the model fits better. All the values of variance are relatively low in the case of ILPD (0.019), indicating that the predictions are stable, but PIDD (0.163) and Diabetes (0.101) have values that are higher, which means that they are sensitive to changes in data. Also, the MCC values show good classification performance with the highest value of Diabetes (0.92), Subsequently, ILPD (0.81) & PIDD (0.77) indicate that the model maintains a commendable equilibrium between correct and inaccurate predictions.

Table 7: Performance Comparison of Recent Studies on the ILPD Dataset

Authors & Year	Protocol	Classifier	Accuracy (%)
Bharadwaj *et al*. (2016)	10-fold Cross Validation	C4.5	69.00
Amare *et al*. (2019)	Train-Test	Naïve Bayes (NB)	71.36
Kabir *et al*. (2019)	—	Stacked Ensemble	73.40
Bihter (2020)	Train-Test	Neural Network	73.28
Sreejith *et al*. (2020)	Train-Test	RF with OSMOTE & without CMVO	82.62
Gan *et al*. (2020)	Train-Test	AdaC-TANBN	68.23
Razali *et al*. (2020)	—	Bayesian Model	70.52
Kuzhippallil *et al*. (2020)	Train-Test	XGBoost and LightGBM	86.00
P. Kumar *et al*. (2021)	10-fold Cross Validation	NWKNN	87.71
Sravani *et al*. (2021)	—	SVM	78.00
K. Gupta *et al*. (2022) [1]	Train-Test	Stacking Ensemble	62.00
Hassim *et al*. (2022) [5]	Train-Test	MLP-BP	70.61

We presented a hard-voting-based ensemble technique, which deals with the difficulties that are associated with tabular diagnostic data in order to enhance both performance and interpretability. Practical diagnostic applications may benefit from the insights uncovered by our work. Results from numerical testing show that our proposed methods either perform as well as, or better than, several well-known methods on the market.

Figure 7 Shows that Deep Forest outperforms the other models in terms of accuracy.

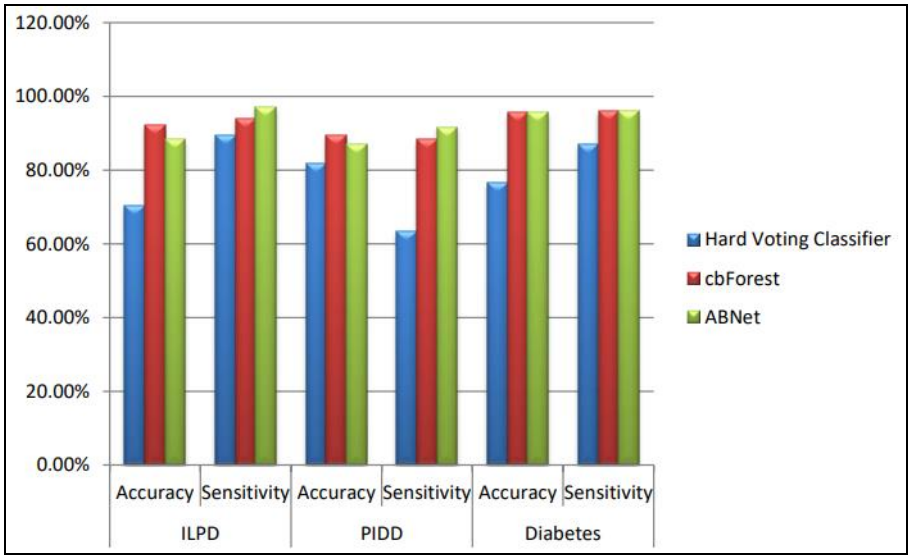

Fig 7: Comparison of the Proposed Models' Performance [%]

In this setting, the significance of deep learning models based on non-neural networks and exploiting non-differential modules becomes apparent. Although it functions similarly to cbForest, ABNet is not nearly as good in terms of accuracy. Despite ABNet's fierce competition, this result suggests that deep ensembles forest cbForest is still the superior method for achieving higher accuracy in our research.

Conclusion

In this paper, a new deep ensemble learning model of medical disease diagnosis is presented that effectively combines preprocessing, feature selection and classification in order to enhance the predictive performance. MICE-based imputation and SOF-based outlier detection methods ensure high-quality data, and the Gini and Permutation Importance methods of finding the best feature subsets. The presented ensemble model, especially the cbForest model, proves to be more accurate and generalizes better than traditional machine learning models based on various datasets. Experimental findings affirm that feature optimization and ensemble learning are very effective in improving classification accuracy as well as preserving model stability. In general, the framework suggested could be a powerful and effective solution to disease prediction and has a high potential to be applied in practice in clinical settings.

References

1. Gupta K, *et al*. Liver disease prediction. In IEEE CSNT, 2022.
2. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: Review of a decade of research. Artificial Intelligence Review,2024:57:273. https://doi.org/10.1007/s10462-024-10884-2
3. Wu Y, *et al*. Imbalanced prediction in epidemiological study: A machine learning approach for stroke prediction. Journal of Biomedical Informatics, 2025.
4. Melnykova N, *et al*. Machine learning for stroke prediction using imbalanced data. Scientific Reports, 2025.
5. Hassim YMM, *et al*. Firefly algorithm neural network learning. Springer, 2022.
6. Olaniyi EO, Adnan K. Diabetes diagnosis using ANN. International Journal of Scientific Engineering,2014:5:754–759.
7. Soltani Z, Jafarian A. ANN approach for diabetes diagnosis. International Journal of Advanced Computer Science,2016:7:89–94.
8. Rakshit S, *et al*. Prediction of diabetes using neural network. Springer, 2017.
9. Ashiquzzaman A, *et al*. Reduction of overfitting in diabetes prediction. Springer, 2018.
10. Tigga NP, Garg S. Prediction of type 2 diabetes. Procedia Computer Science,2020:167:706–716.
11. Jackins V, *et al*. AI-based disease prediction. Journal of Supercomputing,2021:77:5198–5219.
12. Kumari S, Kumar D, Mittal M. Ensemble approach for diabetes classification. International Journal of Cognitive Computing,2021:2:40–46.
13. Barik S, *et al*. Hybrid ML techniques for diabetes prediction. Springer, 2021.
14. Kaur H, Kumari V. Predictive modelling for diabetes. Applied Computing and Informatics,2022:18(1/2):90–100.
15. Sarkar T. XBNet: An extremely boosted neural network. Intelligent Systems with Applications,2022:15:200097.
16. Singh J, Bagga S, Kaur R. Software-based prediction of liver disease with feature selection and classification techniques. Procedia Computer Science,2020:167:1970–1980.
17. Chang V, *et al*. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 2022, 1–17.
18. Caliskan A, Yuksel ME, Badem H, Basturk A. Performance improvement of deep neural network classifiers by a simple training strategy. Engineering Applications of Artificial Intelligence,2018:67:14–23.
19. Zhou ZH, Feng J. Deep forest. National Science Review,2017:6(1):74–86.