Health and Technology (2024) 14:597–613
https://doi.org/10.1007/s12553-024-00835-w

ORIGINAL PAPER

A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets

Palak Mahajan[1] · Shahadat Uddin[2] · Farshid Hajati[3] · Mohammad Ali Moni[4] · Ergun Gide[5]

Received: 16 January 2024 / Accepted: 26 February 2024 / Published online: 27 March 2024

Abstract

Purpose Machine learning models are used to develop and improve various disease prediction systems. Ensemble learning is a machine learning technique that combines many classifiers to increase performance by making more accurate predictions than a single classifier. Although several researchers have employed ensemble techniques for disease prediction, a comprehensive comparative study of these techniques still needs to be provided.

Methods Using 16 disease datasets from Kaggle and the UCI Machine Learning Repository, this study compares the performance of 15 variants of ensemble techniques for disease prediction. The comparison was performed using six performance measures: accuracy, precision, recall, F1 score, AUC (Area Under the receiver operating characteristics Curve) and AUPRC (Area Under the Precision-Recall Curve).

Results Stacking variant of Multi-level stacking showed superior disease prediction performance compared with other bagging and boosting variants, followed by another stacking variant (Classical stacking). Overall, stacking outperformed bagging and boosting for disease prediction. Logit Boost showed the worst performance.

Conclusion The findings of this study can help researchers select an appropriate ensemble approach for future studies focusing on accurate disease prediction.

Keywords Bagging · Boosting · Stacking · Disease prediction · Subvariants · Performance measure

Highlights:

- Our research examines 15 variations of ensemble approaches for disease prediction, providing useful insights into their performance.
- We evaluate the performance of these approaches using six commonly accepted performance measures: accuracy, precision, recall, F1-score, AUC (Area Under the receiver operating characteristics Curve) and AUPRC (Area Under the Precision-Recall Curve).
- Our findings show that stacking variants, notably classical and multi-level stacking, outperform other Bagging and boosting variations in disease prediction.

✉ Shahadat Uddin
shahadat.uddin@sydney.edu.au

[1] Victoria University (Sydney Campus), Sydney, Australia

[2] School of Project Management, Faculty of Engineering, The University of Sydney, New South Wales, Australia

[3] University of New England, Armidale, Australia

[4] AI & Digital Health Technology, Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Bathurst 2795, Australia

[5] Central Queensland University, Sydney, Australia

1 Introduction

Disease diagnosis is a critical step in treating and managing various medical conditions. However, it can be challenging due to the complexity and variability of symptoms and signs. Correct disease diagnosis is essential for effective intervention and patient care [1]. Many scientists have developed machine learning algorithms that accurately identify a broad spectrum of diseases [2–5]. These algorithms can create disease prediction models, enabling early detection and intervention, which are crucial in reducing disease-related mortality [6]. As a result, most medical scientists are drawn to emerging machine learning-based predictive model technologies for disease prediction.

Diabetes, skin disease, kidney disease, liver disease and heart disease are all major chronic diseases that substantially impact health and, if left untreated, can lead to death [7]. Therefore, accurate disease prediction becomes vital in improving patient care and minimising the burden of these chronic conditions. By identifying hidden patterns and relationships in vast healthcare databases, machine learning techniques can assist healthcare professionals in making informed decisions and delivering timely interventions [8]. Ensemble learning is a machine learning technique that aims to improve prediction performance by combining forecasts from several models [1]. Ensemble models reduce the generalisation error in the forecast. The ensemble method reduces model prediction error when the fundamental models are diverse and independent [9].

Bagging, also known as bootstrap aggregating, reduces overfitting and variance by combining predictions from multiple models trained on different subsets of the data [10]. On the other hand, boosting adjusts the weights of misclassified samples iteratively, focusing on difficult-to-classify instances and improving the accuracy of the overall ensemble model. Stacking combines the predictions of multiple models using a meta-learner, which can outperform individual models and other ensemble techniques in various applications [11]. While researchers have used machine learning algorithms extensively for disease prediction, there is a lack of comprehensive studies comparing the performance measures of ensemble learning techniques, such as bagging, boosting, and stacking and their variants, against different significant chronic disease datasets.

A comparative analysis of different ensemble techniques and their variants for disease prediction is crucial in understanding the strengths and limitations of these ensemble approaches. It can help researchers identify the most effective methods for disease prediction [12]. Researchers compared supervised [13, 14] and unsupervised [15] machine learning algorithms for disease prediction. Mahajan et al. [16] conducted a literature review on applying ensemble approaches for disease prediction. However, no study in the current literature compares and contrasts ensemble approaches using multiple datasets. Therefore, the primary objective of this study is to uncover critical trends in disease prediction models based on ensemble learning techniques, specifically bagging, boosting and stacking, and their variants, using performance measures such as accuracy, precision, recall and F1 score. By comparing and evaluating these approaches across various chronic disease datasets, this research provides insights into the effectiveness of different ensemble learning methods for disease prediction.

The datasets used in this study encompass major chronic diseases, including diabetes, chronic kidney disease, liver disease, heart disease, and skin cancer. These diseases were selected due to their prevalence and impact on health outcomes. This study will conduct a comprehensive performance analysis of various ensemble learning techniques by implementing and conducting experiments on 16 machine learning datasets obtained from reputable sources, including Kaggle and the UCI Machine Learning Repository.

For analyses, we considered 15 ensemble algorithms: classical bagging, decision tree, random forest (RF), extra trees, dagging, random subspace, classical boosting, AdaBoost, CatBoost, XGBoost, LightGBM, Logit Boost, Classical stacking, Two-level stacking and Multi-level stacking for disease prediction. Table 1 provides the basic idea, pros and cons for each of these 15 ensemble variants.

2 Materials and methods

2.1 Data source

This study examines 16 datasets from Kaggle and the UCI Machine Learning Repository that are associated with five primary chronic diseases: heart disease, renal disease, liver disease, diabetes and skin cancer. The details of all 16 datasets are provided in Table 2. This table details each dataset's source, number of attributes, total instances, and positive and negative instances. Of these 16 datasets, four are for heart disease, three for liver disease, four for diabetes, three for chronic kidney disease, and two for skin cancer. Data cleaning and preprocessing were performed before conducting the analysis to ensure the quality and integrity of the data. Normalisation was a critical step followed in this procedure since it kept all the data on the same scale and improved the accuracy of the results. While building the model, hyperparameter tuning was performed for all the classifiers to attain better performance.

2.2 Relative performance index

The relative performance index (RPI) is an assessor that collects data results of any performance measure and produces a comparative result for the final assessment [43]. For a given set of performance values, the RPI value is calculated by summing up the difference between each data instance and the minimum value of that dataset. A higher RPI value for an algorithm indicates its superior predictive power compared with other candidates and vice versa [44]. RPI is useful for researchers and practitioners looking to optimise their models for specific datasets. By analysing different variants and calculating their RPI values, it is

Table 1 Basic idea, pros and cons of different ensemble approaches and variants

Algorithm	Basic Idea	Pros	Cons
Bagging	Training multiple models on different data subsets and combining their predictions through averaging or voting	• It reduces overfitting, improves stability, and can handle large datasets [1]. • It is robust against noisy data and parallelisable for efficient computation [2].	• It requires training in multiple models, increases computational cost and may lead to bias [2]. • It may not be effective if base models are highly correlated and have limited interpretability compared to a single decision tree [6, 7].
Decision Tree	A Decision Tree is a tree-based model that works using if-else rules	• Easy to interpret and visualise, handles numerical and categorical data [8]. • It captures nonlinear relationships and is robust against outliers.	• It is susceptible to overfitting and may not capture complex relationships [9]. • It also lacks smooth decision boundaries and handles missing values with additional preprocessing.
Random Forest	Random Forest is a robust decision tree ensemble with random feature selection.	• Capable of handling high-dimensional data, numerical and categorical features [9]. • Provide estimates of feature importance.	• It is slower to train, less interpretable, requires careful tuning, and increases memory usage [10]. • May struggle with datasets with imbalanced class distributions.
Extra Trees	Extra Trees are decision tree construction methods that use randomised feature selection and splitting to improve performance.	• It can handle high-dimensional data and imbalanced datasets well [11]. • Fast training, prediction, and robustness against irrelevant features	• It requires more memory, may be sensitive to noisy data, and may have reduced interpretability [11]. • Randomisation may result in decreased model variance but may also loss essential features.
Dagging	Combines predictions from various models to increase accuracy by lowering model correlation.	• Decorrelated aggregation improves accuracy by reducing correlation between models, increasing robustness [12]. • It can handle complex datasets, reduce overfitting risk, and capture different problem aspects.	• It requires training in multiple models, increasing computational cost, memory usage, and complexity [17]. • Aggregation may not improve performance if individual models are already highly accurate.
Random Subspace	Combining the predictions of various models trained on different feature subsets randomly	• It can handle high-dimensional datasets effectively and be less prone to overfitting [18]. • Parallelisable is so efficient for computation.	• It can be sensitive to noise, increase complexity, require training, and have limited interpretability [18]. • It may not constantly improve performance through random feature selection.
AdaBoost	AdaBoost trains models in a sequence by modifying sample weights.	• This model offers good generalisation and automatically improves weak learners' performance [19]. • It selects essential features and handles imbalanced datasets.	• It is sensitive to outliers, noisy data, and increased computational cost [18]. • It requires careful hyperparameter tuning.
Cat Boost	Algorithm for gradient boosting that supports categorical features	• Effective handling of categorical features provides good performance and accuracy [20]. • Automatically handles missing values and can handle large datasets.	• It can be memory-intensive, slow, require tuning, and increase complexity. • Potentially overfit if hyperparameters are not adequately tuned [20].
LogitBoost	It is a Boosting algorithm that is based on logistic regression.	• Effectively handles complex feature interactions and produces interpretable models [21]. • Handles high-dimensional data and performs well in noise.	• It may require careful tuning of hyperparameters and overfit if weak learners are too complex. • Limited handling of missing values requires additional preprocessing [21].

Table 1 (continued)

Algorithm	Basic Idea	Pros	Cons
XG Boost	This approach gives a scalable gradient-boosting framework	• This model offers good performance, accuracy, regularisation and can handle imbalanced datasets [22]. • It supports parallel processing for faster training and is robust against overfitting [19].	• It has a longer trained time, memory-intensive nature, and requires careful tuning. • Increases complexity and may not perform well with small or sparse datasets [21].
Light GBM	It is a gradient-boosting framework with efficient tree growth.	• This boosting algorithm offers fast training speed and efficient handling of large datasets [22]. • Provides good performance accuracy and handles missing values.	• It is more prone to overfitting and memory intensive. • It requires careful tuning, is less interpretable, and is unsuitable for small datasets [22].
Two-Level Stacking	This stacking approach combines predictions from two levels of models	• It combines strengths from different models and handles high-dimensional datasets [23]. • It provides flexibility in model selection and customisation.	• It requires training in multiple models and increases computational cost. • It suffers complexity while modelling, which may cause potential overfitting [24].
Multi-Level Stacking	This stacking approach combines predictions from a combination of models at different levels.	• Enhance predictive power and flexibility by models [25]. • Can handle complex relationships, capture diverse information, and improve performance.	• It requires complex implementation, which increases computational cost. • Training multiple models at different levels increases interpretation and analysis complexity [23].

possible to identify which ones are most effective for a given task or application, improving the overall quality of the data analysis and decision-making processes. This is the formula for RPI:

$$RPI = \sum_{i=1}^{d} \left(\frac{a_i - a_i^*}{d} \right)$$

where, a_i^* is the minimum value of the list, a_i is the value for the variant under consideration for dataset i, and d is the number of the datasets in the analyses.

2.3 Performance measures

2.3.1 Confusion matrix

A confusion matrix is a method for measuring performance used in statistics and machine learning to evaluate the precision of a classification model [8]. In a confusion matrix, columns correspond to the anticipated class labels, and rows correspond to the true class labels. A confusion matrix is made up of four basic parts (Fig. 1): (a) true positive (TP) is the number of instances that have been correctly predicted as positive from the positive class; (b) true negative (TN) is the number of instances that have been correctly predicted as negative from the negative class; (c) false positive (FP) is the number of instances that have been incorrectly predicted as positive from the negative class; and (d) false negative (FN) is the number of instances that have been incorrectly predicted as negative from the positive class.

Four performance measures considered in this study (i.e., accuracy, precision, recall, and F1 score) are calculated using these confusion matrix values [45]. These metrics can be calculated using the formulas mentioned below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad F1score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

This study also considered two other commonly used performance measures. They are AUC (Area Under the receiver operating characteristics Curve) and AUPRC (Area Under the Precision-Recall Curve). AUC focuses on the trade-off between the true-positive rate (sensitivity) and the false-positive rate, making it appropriate for well-balanced datasets with equally distributed positive and negative examples [42]. AUPRC, on the other hand, focuses more on the precision-recall trade-off, making it appropriate for imbalanced datasets with few positive cases [43].

Table 2 Dataset description

Dataset	Reference	Dataset	Type of Dataset	No. of attributes	No. of instances	Positive/Negative
D1	[26]	Heart Disease Dataset (Comprehensive)	Kaggle	11	1190	629/561
D2	[27]	Health care: Heart attack possibility	Kaggle	13	303	165/138
D3	[28]	Heart Disease Dataset	Kaggle	13	1025	526/499
D4	[29]	Heart Failure Prediction	Kaggle	12	299	96/203
D5	[30]	Liver Disorders	UCI	7	345	200/145
D6	[31]	Indian liver patient dataset	UCI	10	583	167/416
D7	[32]	COVID-19 Effect on Liver Cancer Prediction	UCI	25	450	310/140
D8	[33]	Early-stage diabetes risk prediction dataset	UCI	16	520	320/200
D9	[34]	Diabetes prediction with the KNN algorithm	Kaggle	7	768	268/500
D10	[35]	Diabetes Dataset 2019	Kaggle	17	952	267/685
D11	[36]	Diabetic Retinopathy Debrecen Data Set	UCI	18	1151	611/540
D12	[37]	Chronic Kidney Dataset	Kaggle	25	400	150/250
D13	[38]	Chronic Kidney Disease	Kaggle	13	400	250/150
D14	[39]	Kidney Stone Dataset	Kaggle	7	90	45/45
D15	[40]	Skin Cancer MNIST: HAM10000	Kaggle	6	10015	Multiclass
D16	[41]	Skin Cancer	UCI	35	366	Multiclass

2.4 Experiment setup

The experimental setting for using ensemble approaches to improve binary classification task performance is described in this section. We specifically concentrated on ensemble techniques that use different basic classifiers and hyperparameter tuning techniques for bagging, boosting, and stacking. The intention was to show how various ensemble approaches can be used to increase prediction accuracy. The bagging ensemble methodology combines predictions from several base classifiers to increase performance. The process entails loading a dataset, labelling categorical features, dividing the dataset into test, validation, and training sets, and then instantiating a bagging classifier using a selected base estimator. Libraries must also be imported. GridSearchCV is used for hyperparameter tuning, and the model with the highest accuracy is chosen as the best performer. A thorough classification report is produced after the model has been trained and assessed. A similar procedure was followed for boosting and its variants. Fivefold cross-validation is used to optimise the hyperparameters of each method, improving model performance and offering a thorough evaluation of classification abilities. Both two-level and multi-level stacking are part of the experimental setting for stacking ensemble approaches. Based on the number of levels in the stacking classifier, base classifiers are trained. The predict_proba method is used to produce first-level predictions. GridSearchCV is used once again for hyperparameter tuning. The metamodel produces final predictions, and a classification report is included in the performance evaluation.

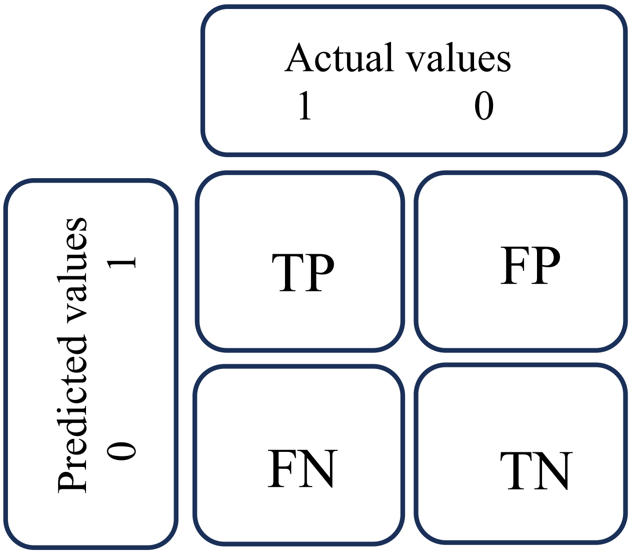

Fig. 1 Confusion matrix

3 Results

3.1 Accuracy comparison

The accuracy outcomes of the ensemble algorithms and their variants are shown in Table 3 against all datasets considered in this study. A bold number in a cell indicates

Table 3 Accuracy (%) of ensemble classifiers and their different variants

Dataset	Type	Bagging						Boosting						Stacking		
		Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classic Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	92%	92%	92%	92%	92%	92%	79%	88%	94%	95%	**96%**	92%	94%	95%	95%
D2		87%	87%	87%	87%	87%	87%	77%	85%	84%	85%	85%	86%	**89%**	74%	84%
D3		98%	98%	98%	98%	98%	99%	66%	91%	99%	99%	99%	94%	99%	**100%**	99%
D4		68%	68%	70%	70%	70%	77%	70%	75%	73%	73%	77%	70%	**99%**	91%	75%
D5	Liver Disease	70%	70%	68%	68%	68%	75%	65%	70%	81%	78%	77%	71%	**82%**	66%	81%
D6		70%	70%	72%	72%	72%	76%	70%	**80%**	74%	72%	70%	77%	76%	71%	72%
D7		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D8	Diabetes	96%	96%	96%	96%	96%	94%	87%	93%	**99%**	97%	98%	97%	98%	97%	**99%**
D9		73%	73%	76%	76%	76%	73%	74%	77%	75%	72%	72%	78%	71%	**80%**	72%
D10		85%	85%	85%	85%	85%	85%	80%	85%	85%	85%	85%	85%	**97%**	85%	85%
D11		69%	69%	**72%**	**72%**	**72%**	71%	66%	68%	69%	**72%**	69%	70%	**72%**	67%	**72%**
D12	Chronic Kidney Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	98%	99%	99%	99%	99%	99%	**100%**	98%	**100%**
D14		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	72%	**100%**	**100%**	95%	**100%**	**100%**	96%	**100%**	**100%**
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	98%	**100%**	**100%**	**100%**
D16		96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	**97%**
Best count		5	5	6	6	6	5	3	5	5	4	5	3	9	6	8

Table 4 Best accuracy frequency and accuracy score against different datasets

Ensemble approach	Datasets																Best count
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging							x				x	x	x	x	x		6
Boosting	x					x	x	x			x	x		x	x		8
Stacking		x	x	x	x		x	x	x	x	x	x	x	x	x	x	14
Best Score	96%	89%	100%	99%	82%	80%	100%	99%	80%	97%	72%	100%	100%	100%	100%	97%	

that the corresponding algorithm (column title) showed the best accuracy performance against the given dataset (row title). Interestingly, datasets D7 and D12 revealed 100% accuracy for all classification algorithms. The last row shows the number of times each algorithm revealed the best performance. Classical stacking (9) has been found to offer the best performance at most times, followed by multi-level stacking (8). Classical boosting and Logit boost performed worst against the same criteria, each revealing the best performance only three times.

Table 4 summarises the outcomes from Table 3 for the three basic ensemble approaches. In doing so, we considered all variants for a basic ensemble technique. For example, we considered all six variants while checking whether bagging produces the best result. If any of them has the best accuracy, we increase the count for the bagging technique. An “x” in a cell designates the ensemble technique that produced the best results for that dataset. For datasets D7, D11, D12, D14, and D15, all three approaches or their variants have shown the best accuracy performance. Again, stacking (14) was the best-performing method, as revealed in the last column.

Apart from the datasets showing the best accuracy for each ensemble technique (i.e., D7, D11, D12, D14 and D15), bagging showed the best accuracy only once (D13), and boosting showed three times (D1, D6 and D8). On the other hand, stacking performed the best nine times (D2-D5, D8-D10, D13 and D16). From this data analysis perspective, it is again stacking that performed best for disease prediction.

3.2 Precision comparison

Table 5 displays the results of precision scores for different ensemble techniques and their variants across disease datasets. All 15 ensemble classifiers considered in this study showed a 100% precision score for datasets D7 and D12. Datasets D12, D13, D15, and D16 consistently performed, giving a precision score of > 90% against each classifier. Regarding how many times a variant reveals the best precision performance (last row of Table 5), Classical Stacking (9) ranked first, followed by Two-level and Multi-level Stacking, each showing the best performance eight times. Classical boosting and logit boost were positioned the lowest in this regard, delivering the best performance four times each. Like the accuracy measure, Classical Boosting and Logit Boosting showed the worst outcome regarding the number of times revealing the best performance. They showed the best performance only four times, much lower than that of classical stacking, which showed the best performance the most times (9).

When variants converged to their corresponding parent ensemble approaches in terms of the number of times revealing the best precision performance, stacking appeared to be the best. The results are presented in Table 6. Stacking showed the best performance 14 times out of 16 datasets, followed by boosting (9) and bagging (8). All variants showed the best precision performance for datasets D7, D8, D12 and D14-D16. For the remaining ten datasets (D1-D6, D9-D11 and D13), stacking achieved the best precision eight times, followed by boosting (3) and bagging (2).

3.3 Recall comparison

For accuracy and precision, the variants of the stacking technique showed the best and second-best performance. Recall outcomes make an exception in this regard – there is a tie for the second-best recall score between random subspace and classical stacking. Each showed the best performance seven times, according to the last row of Table 7. Dataset D12 revealed 100% recall against all ensemble variants. Logit Boost led to the best performance minimum number of times (3) among all variants.

For the three parent ensemble approaches, there is a three-way tie for the best-performing score against datasets D7, D12, D14 and D15, according to Table 8. Stacking scored the best 12 times, followed by boosting (9) and bagging (7). For datasets D3, D7 and D12-D15, stacking showed a 100% recall score.

3.4 F1 score comparison

We observed a similar trend in the F1 score as what we observed for accuracy and precision. Stacking variants outperformed other candidate variants, as detailed in Table 9. Multi-level stacking appeared nine times as the best performer, followed by classical stacking (8) and two-level

Table 5 Precision (%) of ensemble classifiers and their different variants

		Bagging						**Boosting**						**Stacking**		
Dataset	**Type**	*Classical Bagging*	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	92%	92%	92%	92%	92%	92%	73%	88%	94%	**95%**	**95%**	92%	94%	**95%**	**95%**
D2		87%	87%	87%	87%	87%	87%	70%	85%	85%	84%	85%	86%	**90%**	74%	84%
D3		98%	98%	98%	98%	98%	98%	60%	91%	99%	99%	99%	99%	99%	**100%**	99%
D4		67%	67%	68%	68%	68%	77%	66%	80%	76%	73%	77%	70%	**99%**	91%	78%
D5	Liver Disease	80%	80%	75%	75%	75%	76%	66%	69%	81%	78%	77%	71%	**82%**	67%	81%
D6		35%	35%	42%	42%	42%	57%	73%	**84%**	73%	72%	70%	82%	82%	67%	68%
D7		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D8	Diabetes	**99%**	**99%**	**99%**	**99%**	**99%**	96%	92%	93%	**99%**	97%	98%	97%	98%	97%	**99%**
D9		62%	62%	65%	65%	65%	64%	63%	76%	75%	73%	73%	78%	72%	**80%**	72%
D10		82%	82%	82%	82%	82%	80%	69%	82%	82%	82%	82%	82%	**97%**	82%	81%
D11		74%	74%	77%	77%	77%	**78%**	**78%**	67%	70%	73%	70%	70%	72%	67%	73%
D12	Chronic Kidney Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	96%	99%	99%	99%	99%	99%	**100%**	98%	**100%**
D14		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	60%	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**	**100%**	**100%**	98%	**100%**	**100%**	**100%**
D16		**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**
Best count		7	7	7	7	7	7	4	6	6	6	6	4	9	8	8

Table 6 Best precision frequency and precision score against different datasets

Ensemble approach	**Datasets**																**Best count**
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging							x	x			x	x	x	x	x	x	8
Boosting	x					x	x	x			x	x		x	x	x	9
Stacking	x	x	x	x	x		x	x	x	x		x	x	x	x	x	14
Best Score	95%	90%	100%	99%	82%	84%	100%	99%	80%	97%	78%	100%	100%	100%	100%	92%	

Table 7 Recall (%) of ensemble classifiers and their different variants

Dataset	Type	**Bagging**						Boosting						**Stacking**		
		Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	93%	93%	95%	95%	95%	94%	**97%**	88%	94%	95%	95%	92%	94%	95%	95%
D2		87%	87%	87%	87%	87%	87%	**97%**	86%	89%	90%	85%	85%	89%	74%	84%
D3		97%	97%	97%	97%	97%	**100%**	98%	91%	99%	99%	99%	99%	99%	**100%**	99%
D4		48%	48%	52%	52%	52%	44%	86%	75%	73%	73%	77%	70%	**99%**	90%	75%
D5	Liver Disease	67%	67%	71%	71%	71%	**86%**	56%	70%	81%	78%	77%	71%	82%	66%	81%
D6		20%	20%	27%	27%	27%	27%	64%	**79%**	74%	72%	70%	77%	76%	71%	72%
D7		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	87%	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D8	Diabetes	96%	96%	96%	96%	96%	96%	67%	93%	**99%**	97%	98%	97%	98%	97%	**99%**
D9		67%	67%	69%	69%	69%	56%	79%	77%	75%	72%	72%	78%	71%	**80%**	72%
D10		85%	85%	85%	85%	85%	83%	52%	85%	85%	85%	85%	85%	**97%**	85%	84%
D11		67%	67%	69%	69%	69%	64%	62%	68%	69%	**72%**	69%	70%	**72%**	67%	**72%**
D12	Chronic Kidney Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	50%	99%	99%	99%	99%	99%	**100%**	98%	**100%**
D14		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	54%	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**	**100%**	**100%**	98%	**100%**	**100%**	**100%**
D16		96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	**97%**
Best Count		5	5	5	5	5	7	4	5	5	5	4	3	7	6	8

Table 8 Best recall frequency and recall score against different datasets

Ensemble approach	**Datasets**																**Best Count**
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging			x		x		x					x	x	x	x		7
Boosting	x	x				x	x	x			x	x		x	x		9
Stacking			x	x			x	x	x	x	x	x	x	x	x	x	12
Best Score	97%	97%	100%	99%	86%	79%	100%	99%	80%	97%	72%	100%	100%	100%	100%	97%	

Table 9 F1 score (%) of ensemble classifiers and their different variants

Dataset	Type	Bagging						Boosting						Stacking		
		Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	92%	92%	93%	93%	93%	93%	83%	88%	94%	**95%**	**95%**	92%	94%	**95%**	**95%**
D2		87%	87%	87%	87%	87%	87%	82%	86%	89%	**90%**	85%	85%	89%	89%	84%
D3		98%	98%	98%	98%	98%	99%	74%	91%	99%	99%	99%	99%	99%	**100%**	99%
D4		56%	56%	51%	51%	51%	62%	75%	73%	71%	73%	76%	70%	**99%**	94%	73%
D5	Liver Disease	73%	73%	73%	73%	73%	**81%**	61%	68%	**81%**	78%	77%	71%	**81%**	65%	**81%**
D6		25%	25%	33%	33%	33%	36%	69%	**74%**	73%	72%	70%	69%	67%	65%	69%
D7		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D8	Diabetes Disease	97%	97%	97%	97%	97%	96%	90%	93%	**99%**	97%	98%	97%	98%	97%	**99%**
D9		64%	64%	67%	67%	67%	60%	65%	77%	75%	72%	72%	77%	72%	**80%**	72%
D10		82%	79%	79%	79%	79%	76%	74%	82%	82%	82%	82%	82%	**97%**	82%	82%
D11		70%	70%	**73%**	**73%**	**73%**	71%	63%	67%	69%	72%	69%	70%	72%	67%	72%
D12	Chronic Kidney Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	98%	99%	99%	99%	99%	99%	**100%**	98%	**100%**
D14		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	54%	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	96%	**100%**	**100%**	**100%**	**100%**	98%	**100%**	**100%**	**100%**
D16		**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**
Best Count		6	6	7	7	7	6	3	6	7	7	6	4	8	8	9

Table 10 Best F1 score frequency and F1 score against different datasets

Ensemble approach	**Datasets**																**Best Count**
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging							x				x	x	x	x	x	x	7
Boosting	x	x			x	x	x	x				x		x	x	x	10
Stacking	x		x	x	x		x	x	x	x		x	x	x	x	x	13
Best Score	95%	90%	100%	99%	81%	74%	100%	99%	80%	97%	73%	100%	100%	100%	100%	94%	

stacking (8). Datasets D7 and D12 showed a 100% F1 score for all variants. D16 showed the same F1 score (94%) for all variants. Classical boosting appeared minimum times (3) as the best performer.

At the meta-level (i.e., basic ensemble approaches), stacking showed the best F1 score performance 13 times, followed by boosting (10) and bagging (7), according to Table 10. For datasets D7, D12 and D14-D16, all the classifiers have shown the same F1 score.

3.5 AUC comparison

Like in accuracy, precision and F1 score, stacking variants outperformed other candidate variants for AUC, as detailed in Table 11. Multi-level stacking appeared nine times as the best performer, followed by classical stacking (7) and two-level stacking (7). Dataset D15 showed a 100% AUC value for all variants. D16 showed the same AUC score (89%) for all variants. Logit Boost appeared minimum times (3) as the best performer.

According to Table 12, at the meta-level (i.e., basic ensemble approaches), stacking showed the best AUC performance 13 times, followed by boosting (11) and bagging (7). For datasets D2, D7, D12 and D15-D16, all the classifiers showed the same AUC value.

3.6 AUPRC comparison

Multi-level stacking and classical stacking tied in the number of their appearance as the best peformer (8), according to Table 13. Decision tree, XGBoost and two-level stacking appeared six times each as the best performer. Like in AUC, dataset D15 showed a 100% AUPRC score for all variants. Dataset D12 showed the same AUPRC score (98%) for all variants. Classical Boosting and Logit Boost appeared minimum times (3) as the best performer.

According to Table 14, at the meta-level (i.e., basic ensemble approaches), stacking showed the best AUPRC performance 14 times, followed by boosting (9) and bagging (7). For datasets D7-D8, D12-D13 and D15-D16, all the classifiers showed the same AUPRC value.

3.7 Comparing RPI score

Using the results from Table 3, 5, 7, 9, 11 and 13 for 16 datasets, we calculated the RPI score for all performance measures against each variant. Table 15 presents the corresponding RPI score results. Classical stacking showed the highest RPI score for accuracy (11.31%), precision (16.81%) and recall (21.50%) measures. Multi-level stacking showed the highest RPI scores for AUC (9.56%) and AUPRC (12.69%). For the F1 score, Classical Boosting had the highest RPI score (7.06%).

3.8 Comparison of best count statistics

The last rows of Table 3, 5, 7, 9, 11 and 13 show the number of times each variant performed best against accuracy, precision, recall, F1 score, AUC and AUPRC, respectively. Table 16 summarises these six rows to reveal the number of times each variant performed best against all six measures. Stacking variants of multi-level stacking topped the list by appearing 50 times as the best-performing variant. This value is significantly higher than other list values ($p \leq 0.02$) according to the '*inverse normal distribution*' test for a single value. The second highest value was revealed by another stacking variant of classical stacking (48), which is also significantly higher than other remaining values ($p \leq 0.04$). The Logit Boost variant appeared the minimum times (20) as the best performer in this table.

4 Discussion

The ensemble approach, which combines multiple prediction models, proves effective in disease prediction by reducing errors and improving the quality of forecasts. In this study, we evaluated the performance of 15 ensemble techniques, including bagging, boosting, and stacking, using 16 datasets containing information about various diseases. To ensure the reliability of our findings, we rigorously examined how well these ensemble methods performed based on different measures, such as accuracy, precision, recall, and F1 score. We also proceeded to preprocess the data, ensuring it was clean and standardised for accurate predictions.

Our analysis uncovered some interesting trends. For instance, we observed that decision trees performed less effectively in recall and F1 score than other ensemble methods, but bagging demonstrated substantial accuracy and precision. Classical boosting and logit boost performed relatively poorly among the boosting algorithms. However, stacking outperformed other methods, with classical and multi-level stacking exhibiting remarkable results. The repeated success of stacking indicates its reliability and effectiveness as an ensemble method for disease prediction, consistently surpassing other strategies. These findings suggest that stacking could have a meaningful impact on global healthcare by improving disease prediction and management. In addition to these results, our evaluation provides insights into the advantages and limitations of ensemble methods. We observed that ensemble approaches, especially stacking, improve accuracy by reducing outliers. The consistent performance of stacking across diverse datasets highlights its potential as a reliable approach for disease prediction.

Table 11 AUC score (%) of ensemble classifiers and their different variants

Dataset	Type	Bagging						Boosting						Stacking		
		Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	91%	91%	92%	92%	92%	92%	76%	87%	94%	94%	**95%**	91%	85%	**95%**	**95%**
D2		**88%**	85%	**88%**	**88%**	**88%**	**88%**	84%	85%	85%	84%	85%	85%	**88%**	83%	**88%**
D3		98%	97%	99%	99%	99%	99%	78%	91%	98%	98%	98%	94%	82%	98%	**100%**
D4		77%	68%	65%	65%	65%	70%	68%	70%	69%	71%	79%	67%	**88%**	84%	81%
D5	Liver Disease	68%	73%	69%	69%	69%	72%	65%	66%	80%	**82%**	76%	70%	73%	69%	80%
D6		62%	64%	60%	60%	60%	63%	56%	**68%**	65%	63%	58%	60%	56%	55%	66%
D7		96%	96%	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	94%	**99%**	**99%**
D8	Diabetes Disease	83%	84%	85%	85%	85%	85%	80%	89%	**95%**	90%	92%	92%	94%	94%	**95%**
D9		72%	72%	74%	74%	74%	69%	72%	74%	73%	71%	80%	73%	74%	**81%**	75%
D10		83%	84%	85%	85%	85%	85%	80%	85%	87%	85%	86%	85%	**94%**	84%	86%
D11		72%	72%	72%	72%	72%	**77%**	**77%**	67%	70%	69%	72%	**77%**	75%	74%	73%
D12	Chronic Kidney Disease	**98%**	**98%**	**98%**	**98%**	**98%**	94%	**98%**	**98%**	**98%**	**98%**	**98%**	97%	**98%**	**98%**	**98%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	96%	99%	99%	99%	99%	99%	**100%**	97%	**100%**
D14		62%	70%	67%	67%	67%	67%	71%	75%	**79%**	62%	62%	71%	62%	**79%**	78%
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	97%	**100%**	**100%**	**100%**
D16		**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**	**89%**
Best Count		5	4	6	6	6	6	5	5	6	5	5	3	7	7	9

Table 12 Best AUC score frequency and AUC score against different datasets

Ensemble approach	Datasets																Best Count
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging		x					x				x	x	x		x	x	7
Boosting	x	x			x	x	x	x			x	x		x	x	x	11
Stacking	x	x	x	x			x	x	x	x		x	x	x	x	x	13
Best Score	95%	88%	100%	88%	82%	68%	99%	95%	81%	94%	77%	98%	100%	79%	100%	89%	

Table 13 AUPRC score (%) of ensemble classifiers and their different variants

Dataset	Type	Bagging						Boosting						Stacking		
		Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
D1	Heart Disease	89%	89%	90%	90%	90%	89%	73%	84%	91%	**93%**	**93%**	88%	82%	**93%**	**93%**
D2		84%	82%	85%	85%	85%	85%	81%	82%	82%	80%	82%	83%	**85%**	82%	82%
D3		98%	97%	98%	98%	98%	98%	76%	89%	98%	98%	98%	91%	79%	**99%**	98%
D4		66%	58%	54%	54%	54%	65%	62%	64%	60%	59%	64%	55%	61%	64%	**76%**
D5	Liver Disease	71%	76%	72%	72%	72%	74%	67%	70%	80%	**81%**	77%	73%	72%	70%	82%
D6		35%	35%	33%	33%	33%	36%	35%	**50%**	38%	43%	43%	43%	31%	34%	**50%**
D7		**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	87%	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	92%
D8	Diabetes Disease	**86%**	**86%**	**86%**	**86%**	**86%**	71%	65%	**86%**	**86%**	69%	**86%**	**86%**	78%	73%	**86%**
D9		55%	53%	56%	56%	56%	51%	54%	57%	54%	52%	53%	58%	**65%**	59%	62%
D10		75%	66%	63%	63%	63%	68%	66%	68%	67%	69%	77%	65%	**86%**	82%	79%
D11		71%	70%	70%	70%	70%	**72%**	**72%**	66%	69%	**72%**	68%	68%	69%	71%	70%
D12	Chronic Kidney Disease	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**
D13		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	94%	99%	**100%**	99%	99%	99%	**100%**	98%	**100%**
D14		40%	49%	43%	43%	43%	43%	49%	52%	58%	42%	42%	49%	41%	56%	**68%**
D15	Skin Disease	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	99%	**100%**	**100%**	**100%**
D16		68%	**73%**	69%	69%	69%	72%	65%	66%	69%	66%	65%	65%	**73%**	**73%**	76%
Best Count		5	6	5	5	5	5	3	5	5	6	5	3	8	6	8

Table 14 Best AUPRC score frequency and AUPRC score against different datasets

Ensemble approach	Datasets																Best Count
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*	*D15*	*D16*	
Bagging							x	x			x	x	x		x	x	7
Boosting	x				x	x	x	x			x	x	x		x		9
Stacking	x	x	x	x		x	x	x	x	x		x	x	x	x	x	14
Best Score	93%	85%	99%	76%	81%	50%	94%	86%	65%	86%	72%	98%	100%	68%	100%	73%	

Table 15 RPI score for ensemble classifiers and their variants

Ensemble approach	**Bagging**						**Boosting**						**Stacking**		
	Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
Accuracy	7.31%	7.29%	7.81%	7.81%	7.81%	8.66%	0.61%	7.42%	8.83%	8.16%	8.46%	7.81%	**11.31%**	8.21%	8.89%
Precision	10.25%	10.25%	10.81%	10.81%	10.81%	12.06%	3.13%	12.63%	13.81%	13.31%	13.31%	13.25%	**16.81%**	12.88%	13.63%
Recall	14.38%	14.38%	15.69%	15.69%	15.69%	15.00%	9.25%	19.63%	21.25%	20.94%	20.56%	20.25%	**23.50%**	20.38%	21.06%
F1 score	3.30%	3.49%	2.88%	2.88%	2.88%	2.27%	**7.06%**	0.06%	2.13%	2.00%	1.56%	0.75%	4.19%	2.19%	1.81%
AUC	5.56%	5.81%	5.75%	5.75%	5.75%	6.19%	2.44%	5.75%	8.13%	6.50%	7.38%	6.00%	6.38%	8.06%	**9.56%**
AUPRC	7.56%	7.31%	6.38%	6.38%	6.38%	6.69%	2.19%	7.25%	8.44%	6.63%	8.13%	6.56%	6.56%	8.56%	**12.69%**

Table 16 Comparison of approaches considering the number of times showing the best performance

Performance measure	**Bagging**						**Boosting**						**Stacking**		
	Classical Bagging	*Decision Tree*	*Dagging*	*Random Forest*	*Extra Tree*	*Random Subspace*	*Classical Boosting*	*AdaBoost*	*Cat Boost*	*XGBoost*	*Light GBM*	*Logit Boost*	*Classical Stacking*	*Two-Level*	*Multi-Level*
Accuracy	5	5	6	6	6	5	3	5	5	4	5	3	9	6	8
Precision	7	7	7	7	7	7	4	6	6	6	6	4	9	8	8
Recall	5	5	5	5	5	7	4	5	5	5	4	3	7	6	8
F1 score	6	6	7	7	7	6	3	6	7	7	6	4	8	8	9
AUC	5	4	6	6	6	6	5	5	6	5	5	3	7	7	9
AUPRC	5	6	5	5	5	5	3	5	5	6	5	3	8	6	8
Count	33	33	36	36	36	36	22	32	34	33	31	20	48	41	**50**

Although this research considered 16 benchmark research datasets from two highly regarded open-access data repositories, these datasets may not capture the full complexity and variability of other real-world health data, such as the data from clinical settings. Moreover, most of these datasets are highly balanced. Clinical settings often encounter imbalanced data [44]. This limitation of our study opens a new research scope for the future, establishing research collaborations with healthcare providers that have access to such data to validate the findings of this study. Integrating any computation models, including the best one this study observed (i.e., classical and multi-level stacking), with the present healthcare environments is challenging, primarily due to computational complexity [45] and cost-effectiveness [46]. While this study focused on the robustness of the theoretical findings, it is crucial to research the potentiality of their real-world applications. Numerous studies [e.g., 47] highlighted the importance of adopting advanced technologies and computational models appropriately in healthcare settings. Our research echoed this importance once again. Overfitting could be another limitation of this study. Despite their effectiveness, ensemble approaches are sometimes prone to overfitting [48], especially when working with intricate models like stacking or boosting. Our consideration of GridSearchCV for hyperparameter tuning that maximises performance while minimising overfitting and cross-validation helps reduce the negative impact of this overfitting issue.

Our findings could potentially add new thoughts to improving ensemble model performance. These findings offer several directions for future research. Comparative analyses can help determine which ensemble strategy is most suitable for healthcare scenarios. Fine-tuning the methods and optimising individual algorithms can further enhance prediction accuracy. Furthermore, exploring specialised feature engineering techniques for specific domains may improve the predictive power of ensemble models. Real-world validation is essential to test their performance in healthcare settings to ensure the practical application of ensemble models. When using ensemble models for illness prediction, ethical considerations are critical. Future studies should focus on protecting privacy, minimising discrimination based on projected health effects, and ensuring responsible and equitable use. Integrating ensemble models with existing medical technologies holds promise for improving disease prediction accuracy and usefulness, ultimately benefiting patients and healthcare providers. Finally, our study highlights the strengths and potential of ensemble methods in disease prediction, with stacking emerging as a standout performer. Our recommendations for future research encompass comparative analysis, algorithm refinement, interpretability, validation, ethical considerations, and seamless integration with other healthcare technologies. These research avenues promise to advance ensemble approaches for disease prediction, leading to more accurate predictions and improved healthcare outcomes.

5 Conclusion

In this research, we evaluated the performance of various algorithms and their variations in the context of disease prediction through ensemble techniques. The findings consistently favoured the stacking technique over other ensemble strategies, revealing its effectiveness in accurately predicting diseases across diverse datasets. Notably, stacking achieved 100% accuracy on some datasets, highlighting its potential as a robust and reliable ensemble method. While bagging classifiers such as Dagging, Random Forest, Extra Trees, and Random Subspace demonstrated strong performance within the bagging ensemble models, stacking outperformed individual techniques such as CatBoost, XGBoost, and LightGBM in the boosting category. Classical boosting and LogitBoost emerged as the weakest classifiers among the various ensemble approaches assessed.

These results provide valuable guidance for selecting the most suitable algorithm for disease prediction. Notably, stacking, particularly the classical stacking and multi-level stacking algorithms, emerged as the most reliable and precise ensemble methods, outperforming other approaches across all performance metrics, showing the advantage of combining the strengths of multiple models and reducing bias and variance in predictions. The implications of these findings are significant for the field of disease prediction, as they enable healthcare professionals to enhance the accuracy of disease prediction models, potentially leading to earlier diagnosis, expedited treatment, and improved patient outcomes. Further research is warranted to explore aspects such as interpretability, optimisation, ethical considerations, and the integration of ensemble models with other medical technologies. Addressing these aspects can advance the field, resulting in more accurate and reliable predictive models for disease prediction.

Author contributions PM: Analysis, Coding, Writing. SU: Conceptualisation, Supervision, Analysis and Writing. FH, MAM and EG: Writing.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability All data used in this study are publicly available in the Kaggle and UCI Machine Learning Repository.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access. 2022;10:99129–49.
2. Ramesh D, Katheria YS. Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach. Health Technol. 2019;9:533–45.
3. Lu H, Uddin S. Embedding-based link predictions to explore latent comorbidity of chronic diseases. Health Inform Sci Syst. 2022;11(1):2.
4. Uddin S, Wang S, Lu H, Khan A, Hajati F, Khushi M. Comorbidity and multimorbidity prediction of major chronic diseases using machine learning and network analytics. Expert Syst Appl. 2022;205: 117761.
5. Hossain ME, Khan A, Uddin S. Understanding the comorbidity of multiple chronic diseases using a network approach. In Proc Austral Comput Sci Week Multiconference. 2019;1–7.
6. Nikookar E, Naderi E. Hybrid ensemble framework for heart disease detection and prediction. Int J Adv Comput Sci Appl. 2018;9(5):243–8.
7. Igodan EC, Thompson AF-B, Obe O, Owolafe O. Erythemato squamous disease prediction using ensemble multi-feature selection approach. Int J Comput Sci Inf Secur. 2022;20:95–106.
8. Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular disease detection using ensemble learning. Comput Intell Neurosci. 2022;2022:9.
9. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, Nappi M. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access. 2021;9:39707–16.
10. Chaurasia V, Pandey MK, Pal S. Chronic kidney disease: a prediction and comparison of ensemble and basic classifiers performance. Human-Intelligent Syst Integr. 2022;4(1–2):1–10.
11. Zubair Hasan K, Hasan Z. Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease. In: Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, vol. 1. Springer; 2019. pp. 415–26.
12. Yariyan P, Janizadeh S, Van Phong T, Nguyen HD, Costache R, Van Le H, Pham BT, Pradhan B, Tiefenbacher JP. Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. Water Resour Manage. 2020;34:3037–53.
13. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inf Decis Mak. 2019;19(1):1–16.
14. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12(1):1–11.
15. Lu H, Uddin S. Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets. Health Technol. 2024;14(1):141–54.
16. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. Healthcare. 2023;11(12):1808.
17. Kotsianti S, Kanellopoulos D. Combining bagging, boosting and dagging for classification problems. In Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12–14, 2007. Proceedings, Part II 11. 2007. Springer.
18. Basar MD, Akan A. Detection of chronic kidney disease by using ensemble classifiers. In 2017 10th International Conference on Electrical and Electronics Engineering (ELECO). IEEE; 2017. pp. 544–47.
19. Shorewala V. Early detection of coronary heart disease using ensemble techniques. Inf Med Unlocked. 2021;26:100655.
20. Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, Yu J, Li C, Yu F, Ren Z. Machine learning models for data-driven prediction of diabetes by lifestyle type. Int J Environ Res Public Health. 2022;19(22):15027.
21. Nahar N, Ara F, Neloy MAI, Barua V, Hossain MS, Andersson K. A comparative analysis of the ensemble method for liver disease prediction. In 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET). IEEE; 2019. pp. 1–6.
22. Singh V, Gourisaria MK, Das H. Performance analysis of machine learning algorithms for prediction of liver disease. In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). IEEE; 2021. pp. 1–7.
23. Liza FR, Samsuzzaman M, Azim R, Mahmud MZ, Bepery C, Masud MA, Taha B. An ensemble approach of supervised learning algorithms and artificial neural network for early prediction of diabetes. In 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE; 2021. pp. 1–6.
24. Abdollahi J, Nouri-Moghaddam B. Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran J Comput Sci. 2022;5:205–20.
25. Kuzhippallil MA, Joseph C, Kannan A. Comparative analysis of machine learning techniques for indian liver disease patients. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE; 2020. pp. 778–82.
26. Alizadehsani R, Roshanzamir M, Abdar M, Beykikhoshk A, Khosravi A, Panahiazar M, Koohestani A, Khozeimeh F, Nahavandi S, Sarrafzadegan N. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Sci data. 2019;6(1):227.
27. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart disease UCI mach learn repository. 2020. https://doi.org/10.24432/C52P4X.
28. Lapp D. Heart disease dataset. 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.
29. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inf Decis Mak. 2020;20(1):1–16.
30. Forsyth RS. Liver disorders data set. 1990. https://archive.ics.uci.edu/ml/datasets/Liver+Disorders.
31. Ramana BV. Indian liver patient dataset data set. 2012. https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29.

32. Fedesoriano. COVID-19 effect on liver cancer prediction dataset. 2022. Available from: https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset.
33. Early stage diabetes risk prediction dataset. 2020. Available from: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset.
34. Mahgoub A. Diabetes prediction system with KNN algorithm. 2021. https://www.kaggle.com/abdallamahgoub/diabetes .
35. Tigga NP. Diabetes Dataset 2019. 2020. Available from: https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019.
36. Antal B, Hajdu A. An ensemble-based system for automatic screening of diabetic retinopathy. Knowl Based Syst. 2014;60:20–7.
37. Iqbal M. Chronic kidney disease dataset. 2017. https://www.kaggle.com/datasets/mansoordaku/ckdisease.
38. Pandit AK. Chronic kidney disease. 2020. Available from: https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease.
39. Ghadiya H. Kidney stone dataset. Available from: https://www.kaggle.com/datasets/harshghadiya/kidneystone.
40. Mader S, Skin Cancer MNIST. : HAM10000. 2018. Available from: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000.
41. Ilter N. Dermatology data set. 1998. https://archive.ics.uci.edu/ml/datasets/Dermatology.
42. de Hond AA, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health. 2022;4(12):e853-855.
43. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68(8):855–9.
44. Tarekegn AN, Giacobini M, Michalak K. A review of methods for imbalanced multi-label classification. Pattern Recogn. 2021;118:107965.
45. Chen P-T, Lin C-L, Wu W-N. Big data management in healthcare: adoption challenges and implications. Int J Inf Manag. 2020;53:102078.
46. Lokkerbol J, Adema D, Cuijpers P, Reynolds CF III, Schulz R, Weehuizen R, Smit F. Improving the cost-effectiveness of a healthcare system for depressive disorders by implementing telemedicine: a health economic modeling study. Am J Geriatric Psychiatry. 2014;22(3):253–62.
47. Colicchio TK, Facelli JC, Del Fiol G, Scammon DL, Bowes WA III, Narus SP. Health information technology adoption: understanding research protocols and outcome measurements for IT interventions in health care. J Biomed Inform. 2016;63:33–44.
48. Grushka-Cockayne Y, Jose VRR, Lichtendahl Jr KC. Ensembles of overfit and overconfident forecasts. Manage Sci. 2017;63(4):1110–30.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.