scientific reports

Check for updates

OPEN

Advancing ischemic stroke diagnosis and clinical outcome prediction using improved ensemble techniques in DSC-PWI radiomics

Mazen M. Yassin[1,2,3], Jiaxi Lu[2,4], Asim Zaman[1,2,4], Huihui Yang[2,4], Anbo Cao[2,4], Xueqiang Zeng[2,4], Haseeb Hassan[2], Taiyu Han[2,4], Xiaoqiang Miao[2,5], Yongkang Shi[2,4], Yingwei Guo[6], Yu Luo[7] & Yan Kang[1,2,5,8]✉

Ischemic stroke is a leading global cause of death and disability and is expected to rise in the future. The present diagnostic techniques, like CT and MRI, have some limitations in distinguishing acute from chronic ischemia and in early ischemia detection. This study investigates the function of ensemble models based on the dynamic radiomics features (DRF) from the dynamic susceptibility contrast perfusion-weighted imaging (DSC-PWI) ischemic stroke diagnosis, neurological impairment assessment, and modified Rankin Scale (mRS) outcome prediction). DRF is extracted from the 3D images, features are selected, and dimensionality is reduced. After that, ensemble models are applied. Two model structures were developed: a voting classifier with 6 bagging classifiers and a stacking classifier based on 4 bagging classifiers. The ensemble models were evaluated on three core tasks. The Stacking_ens_LR model performed best for ischemic stroke detection, the LR Bagging model for NIH Stroke Scale (NIHSS) prediction, and the NB Bagging model for outcome prediction. These outcomes illustrate the strength of ensemble models. The work showcases the role of ensemble models and DRF in the stroke management process.

Keywords Stroke classification, NIHSS prediction, Dynamic radiomics features, DSC-PWI.

Ischemic stroke is a major global health problem since it ranks second among the leading causes of death and disability due to cerebrovascular diseases around the world. As pointed out, over the last three decades from 1990 to 2019, a significant number of deaths were caused by ischemic stroke, and the number is projected to grow even further by 2030[1,2]. Ischemic stroke, which accounts for up to 80% of all strokes, occurs due to a sudden reduction of blood flow to a specific brain area, resulting in tissue damage and neurological disorders[3,4]. Smoking, obesity, high blood pressure, and increased cholesterol are the major risk factors of ischemic stroke[5]. Thrombolysis is the main technique for ischemic stroke treatment that, however, results in oxygen increase and free radical production over the brain, leading to an aggravation of brain injury[6].

Current diagnostic methods for ischemic stroke detection, such as non-contrast Computed Tomography (CT) as well as Magnetic Resonance Imaging (MRI), have problems of differentiation between the acute and the chronic infarcts and the detection of the early signs of ischemia[7,8]. To deal with these constraints, innovative approaches have been developed. For example, a hybrid approach with the combination of Convolutional Neural Network (CNN), and Kernel K-Means clustering demonstrate a great performance in recognizing ischemic stroke cases from MRI images and achieving high accuracy, specificity, precision, and F1-score[8]. Moreover,

[1]School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen 518055, China. [2]College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China. [3]Biomedical Engineering Department, Faculty of Engineering, Minia University, Menia 61111, Egypt. [4]School of Applied Technology, Shenzhen University, Shenzhen 518055, China. [5]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China. [6]School of Electrical and Information Engineering, Northeast Petroleum University, Daqing 163318, China. [7]Department of Radiology, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, Shanghai 200434, China. [8]Faculty of Data Science, City University of Macau, Macau, China. ✉email: kangyan@sztu.edu.cn

the Alberta Stroke Program Early CT Score (ASPECTS) software automated version has proved to be fairly accurate in early ischemic CT changes, with a slice thickness of 5 mm being the best for adequate results[9]. These innovations underline the prospect of more efficient and exact diagnosis of ischemic stroke, thereby assisting in the quick and accurate management of patients.

Furthermore, dynamic susceptibility contrast perfusion-weighted imaging (DSC-PWI) is a functional MRI technique that can be used to evaluate cerebral perfusion impairment in cerebral tissues especially in cases of ischemic stroke. Recent research shows the opportunity of DSC-PWI to demonstrate the areas of reduced perfusion and to deliver the important information about the blood circulation state needed for the acute stroke management. The cerebral blood flow (CBF) and the cerebral blood volume (CBV) maps derived from the DSC-PWI processing are accurate for the identification of hypoperfusion areas suggestive of ischemia[10]. This technique has been shown to estimate final infarct size soon after ischemic stroke onset, thus it can be applied non-invasively in the clinical and experimental practice. When comparing DSC-PWI with arterial spin labeling (ASL), the studies have shown that both methods are similarly useful for evaluating ischemic penumbra and infarct core, although DSC-PWI provides more quantitative data and is therefore more helpful in making a diagnosis[11]. Also, there is nearly perfect concordance between DSC-PWI and T2*GRE imaging in hemorrhages and acute hemorrhagic transformation in the ischemic stroke patients; this can replace conventional methods and hasten treatment[12]. DSC-PWI plays a crucial role in decision-making processes in treating the diseases as it can differentiate between the healthy and infarcted brain tissue, as well as evaluate the perfusion states and collateral circulation and overall enhances the clinical management and patients' outcomes[13].

The National Institutes of Health Stroke Scale (NIHSS) represents a vital measurement tool for assessing the degree of neurological impairment in stroke patients[14], including the performance of consciousness, motor function, sensation, and language with scores ranging from 0 to 42[15]. Comparing the NIHSS score with other factors, age, hypertension, and collateral grading can provide an idea of how stroke patients will recover and how the treatment plan will be decided[14]. Consequently, the NIHSS score is an essential aspect of determining the level of neurological dysfunction and providing individualized stroke treatments that affect the patient's rehabilitation prospects[16]. Using medical imaging can improve stroke diagnostics but there are situations where patients have low NIHSS scores and still have strokes which indicate limitations of depending only upon NIHSS or imaging for the diagnosis. Incorporating imaging features that correlate with the NIHSS scores alongside imaging techniques of stroke lesions can enhance diagnostic accuracy, especially in cases when a conventional diagnostic procedure might not reveal a stroke, thus, making it possible to determine the severity and prognosis more accurately[17]. Therefore, this integration of tests could potentially produce more accurate diagnoses and better outcomes by giving a more integral understanding of the stroke's extent.

The significance of the modified Rankin Scale (mRS) prediction for personalized stroke rehabilitation planning is emphasized by numerous studies, which confirm that refined predictive models contribute to the improvement of rehabilitation outcomes. Machine learning has proved as an efficient way to predict functional recovery and clinical outcomes in mRS according to studies[18,19]. Such methods help to implement tailored rehabilitation programs that are accurate to specific recovery profiles of stroke patients, thus improving their quality of life. The development of standardized instruments that can be used to integrate clinical data will help to achieve the precision of rehabilitation prognoses and to equalize the provision of discharge and rehabilitation services[20]. In addition, some studies considered how the lesion topography and mRS-based models could be used to adjust the acute interventions and the resource allocation to design personalized care strategies for the patients[21,22]. Furthermore, other studies have been instrumental in the field through the development of predictive tools and intensive rehabilitation strategies that have greatly contributed to better mRS outcomes[23,24]. This has helped in early intervention and customized rehabilitation planning.

Radiomics, a burgeoning field in medical imaging, extends its applications beyond oncology into radiology and ophthalmology, showcasing its versatility in various medical domains. In radiology, radiomics aids in screening, disease detection, diagnosis, staging, and prognosis, as well as in finding and predicting biological correlates[25–27]. This is supported by research highlighting the use of radiomics in cardiovascular imaging, specifically in ischemic heart disease (IHD)[28]. Moreover, in ophthalmology, Studies have shown the effectiveness of radiomics in various ophthalmic conditions like diabetes mellitus (DM), diabetic retinopathy (DR), referable DR (R-DR), and dysthyroid optic neuropathy (DON)[29–32]. The integration of radiomics into these fields offers a promising avenue for enhancing clinical decision-making and personalized patient care, emphasizing the broad spectrum of applications beyond oncology[33]. The high-dimensional data obtained from radiomics can acknowledge individualized medicine by identifying the therapy responses as well as the precise clinical outcomes, that is, personalized healthcare. Thus, radiomics represents an essential tool in recognizing the pathophysiology and heterogeneity of diseases such as ischemic stroke because it provides useful diagnostic information from medical images, with precision[34,35].

Ensemble models provide the foremost benefits in comparison to other models in machine learning for ischemic stroke diagnostic imaging and outcome prediction[36]. They can potentially increase reliability by combining multiple models, which may enhance overall classification accuracy and robustness, depending on factors such as model correlation and dataset size[34,37]. The multi-model ensemble approach, such as the OEDL (optimized ensemble of deep learning), is superior for predicting stroke prognosis than single models[38]. Additionally, different structures of ensemble models such as the voting classifier that combines Support Vector Machine (SVM), Random Forest, and Decision Tree (DT) classifiers are effective in predicting stroke diagnoses in real time by using Electrocardiography (ECG) and Photoplethysmography (PPG) data[39]. The hybridization of multi-sequence MRI using ensemble methods has shown outstanding potential for the diagnosis of malignant soft tissue tumors[40], which proves that ensemble models are well-suited for radiomics purposes[41]. Ensemble models in radiomics for ischemic stroke offer superior predictive power by combining features from DWI and

T2-FLAIR sequences[42] and combining MRI radiomics and clinical data[43] for accurate diagnosis and outcome prediction[44].

This study aims to show the role of ensemble models based on the certain dynamic radiomics features (DRF) of DSC-PWI in diagnosing ischemic stroke, assessing neurological impairments, and predicting outcomes (90-day mRS). The main contributions will be outlined in three key areas. Firstly, the paper focuses on the search for the role of ensemble models on DRF in ischemic stroke. Initially, the radiomics features extracted from 3D images within the time-segregated DSC-PWI are used to obtain the DRF for the whole brain. Subsequently, feature selection and dimensionality reduction techniques are used. Applying ensemble models to DRF has been found to be clinically relevant for stroke diagnosis, NIHSS assessment, and outcome prediction, making it a promising tool for clinical application. The second issue that the study emphasizes on building two model structures, the first model is bagging classifiers with voting classifiers and the second model is stacking classifier based on bagging classifiers. In the final part, the study utilizes ensemble models with the aim of reducing variance and bias in predictive models, mitigating the impact of outliers in the datasets, and improving the generalization of the predictive models. Therefore, this study can give us useful information about the ensemble models using DRF for overall stroke management and outcome prediction.

Materials and methods

Detailed materials and methods are introduced in the following subsections. The materials are described in "Materials" sect., and the methods are shown in "Methods".

Materials

IRBs at the Shanghai Fourth People's Hospital, which is an affiliate of the Tongji University School of Medicine, approved this retrospective study and waived the requirement for informed consent (Approval Code: 20200066-01; Approval Date, 15 May 2020). All experiments were performed in accordance with relevant guidelines and regulations. The datasets in our study were extracted from the neurology department of the Shanghai Fourth People's Hospital, affiliated with the Tongji University School of Medicine, China, in the period from 2013 to 2016. A total of 156 DSC-PWI images from 88 patients were retrospectively assessed and included. All patients were imaged within 24 h of their symptom onset. The DSC-PWI screening was conducted at least twice, once in the pretreatment phase and once in the post-treatment phase, for 68 patients. A total of 78 (50%) DSC-PWI images were diagnosed with ischemic stroke based on the results of clinical examination. The key clinical information is 90-day mRS and outcome NIHSS. The DSC-PWI image was scanned on a 1.5T Siemens Avanto MR scanner, primarily utilizing T2*-weighted sequences. None of the patients had metal implants or calcifications, which can affect DSC imaging. To ensure robust evaluation of our models, we divided the dataset into training and test sets. The dataset was split into 80% for training and 20% for testing. Table 1 provides the details.

In this study, the DSC-PWI dataset was preprocessed as per the methods presented in[45], including registration and voxel-wise smoothing with a 1×3 kernel triple moving average filter to remove the noise and the positional deviation, followed by delineation of brain tissue using the software package FMRIB Software Library (FSL)[46]. This dataset was used to compute DRF features, which were extracted from the segmented 3D images using the PyRadiomics package (version 3.0.1) to obtain very detailed feature groups like First_order and Gray Level Co-occurrence Matrix (GLCM)[45]. Using T-test analysis, the significant DRF is extracted according to classification output categories. To establish the basis of fact for evaluating the diagnostic and prognostic usefulness of these features in ischemic stroke it was necessary to locate and mark (1 for presence, 0 for absence) ischemic lesions by using Rapid Processing of Perfusion and Diffusion (RAPID) software[47] and to have an assessment of the level of impairment by the NIHSS scores and an estimation of outcome prediction by the mRS scores at 90 days. In Table 2, the preprocessing steps, radiomic features extraction, and Ground Truth data are summarized in detail. The ground truth equal to 1 means a patient with ischemic stroke lesions, neurological impairment (more than zero score), or poor outcome (the 90-day mRS more than 1), and the ground truth equal to zero means no ischemic stroke lesions, normal neurological function, or good outcome. Table 2 provides detailed information about the preprocessed data used in this study. The "Ground Truth" section of the table presents the counts of each label for the three classification tasks. The "DRF Details" section of the table outlines the number of each type of DRF extracted for the three classification tasks. These features are categorized by their type.

Data set Information		Scanning Parameters of DSC-PWI Images	
Numbers of patients	88	TE/TR	$32/1590$ ms
Datasets (sets)	156	Matrix	256×256
Female (%)	39 (25%)	FOV	230×230 mm^2
Age (Mean $\pm$ Std)	69.919 ± 6.747 years	Thickness	5 mm
NIHSS (Mean $\pm$ Std)	6.275 ± 6.875	Number of measurements	50
90-day mRS	2.60 ± 2.34	Spacing between slices	6.5 mm
Ischemic stroke (%)	78 (50%)	Pixel bandwidth	1347 Hz/ pixel
		Number of slices	20

Table 1. Data set details.

Ground Truth (count)			
Label	Ischemic Stroke	NIHSS	90-Day mRS (Outcome prediction)
1	78	61	55
0	78	95	101

DRF Details (number of features)							
Output	First_order	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Stroke Detection	5118	7698	3800	4117	3737	1352	25,822
NIHSS prediction	2061	2655	866	1016	1289	437	8324
Outcome Prediction	2089	2650	1304	1254	1439	467	9203

Table 2. Preprocessed data details. First-order statistics (First_order), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependency Matrix (GLDM) and Neighboring Gray-Tone Difference Matrix (NGTDM).

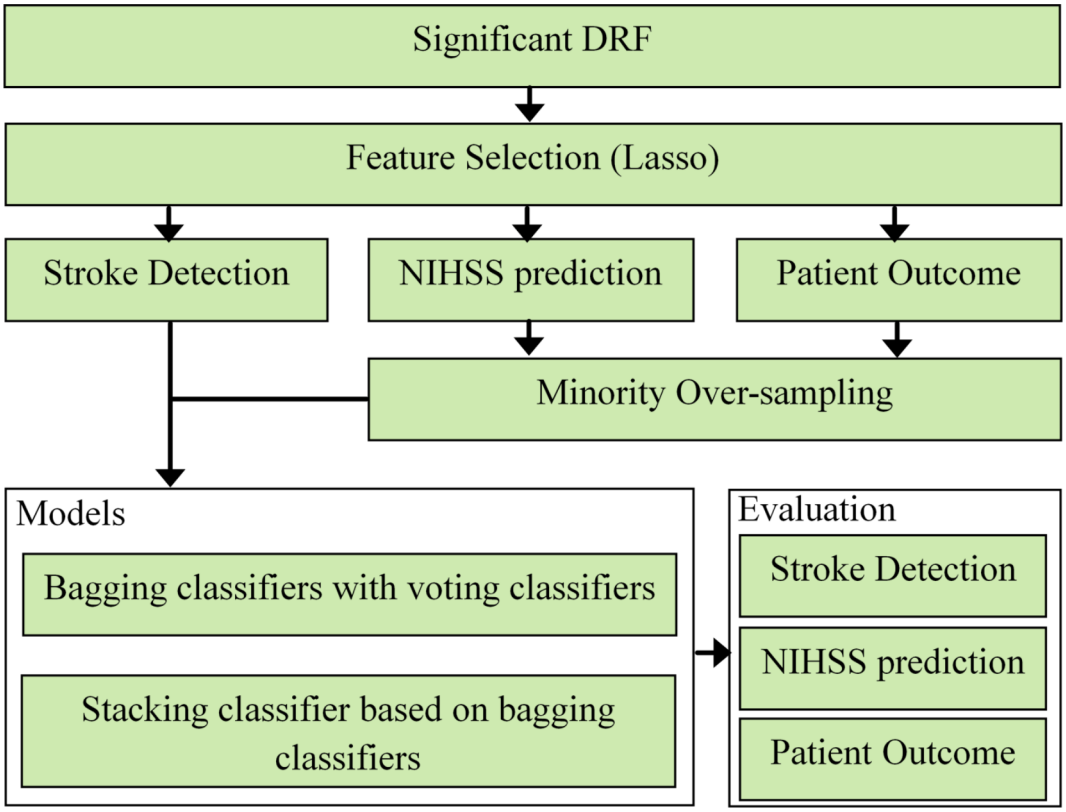

Fig. 1. The flowchart of the proposed method in this study.

Methods

This study proposes the method in many steps: feature selection, processing the classes imbalance, models' structure then evaluation as shown in Fig. 1.

Feature selection

Feature selection is the key component of machine learning algorithms and feature processing is required for better algorithm performance. Lasso algorithm is considered one of the most potent ways of feature selection with target variables[45,48]. As a result, this study has chosen the DRF that have non-zero coefficients by using Lasso depending on the three sets of ground truth, which are the outstanding DRF. The Lasso was implemented by the LassoCV function imported from the sklearn.linear_model package in Python 3.12 and the cv was set as 10 in the function. The 'cv' variable refers to the number of folds used in cross-validation, which is a technique to evaluate the performance of the model by dividing the dataset into multiple subsets. The choice of 10 folds is a commonly used practice as it provides a good balance between bias and variance, ensuring a reliable estimation of the model's performance. The mathematical equation of Lasso is depicted in Eq. (1).

$$Lasso\left(F_{t-test}, K\right) = \min_{\beta_k} \left\{ \frac{1}{2n_k} \sum_{i=1}^{n_k} \left(y_{ik} - \beta_{0k} - \sum_{j=1}^{p_k} \beta_{jk} x_{ijk} \right)^2 + \lambda_k \sum_{j=1}^{p_k} \left| \beta_{jk} \right| \right\} \quad (1)$$

where the task is k; $Lasso(F_{t-test}, K)$ represents the selected outstanding DRF for the evaluation task from the significant DRF; y_{ik} represents the ground truth for the i-th observation; x_{ijk} denotes j-th the independent DRF for the i-th observation; β_{0k} is the intercept term; β_{jk} are the coefficients for the independent DRF ; n_k is the number of observations, and p_k is the number of the independent DRF for the k-th task; λ_k is the regularization parameter.

Minority over-sampling

During our study, we found the distribution of labels for NIHSS prediction and patient outcome tasks was imbalanced largely with minority classes underrepresented. As a solution to this problem and to make our predictive models more robust, we used the Synthetic Minority Over-sampling Technique (SMOTE) impact on our training data[49,50]. SMOTE aims at extrapolating new instances from the minority class. Therefore, the class distribution is balanced, and information loss is minimized which is the case in under-sampling methods. The mathematical equation of SMOTE is depicted in Eq. (2).

$$x_{\text{new}} = x_i + \lambda \cdot (x_z - x_i) \quad (2)$$

Where x_{new} is the synthetic sample; x_i is a randomly chosen minority class sample; x_z is one of the k nearest neighbors of x_i (also belonging to the minority class); λ is a random number between 0 and 1.

The formulation of this technique allows for the creation of more natural and diverse new examples by generating examples between the existing examples of the minority class. The synthetic examples generated by SMOTE complement real data, creating a more balanced dataset, which can potentially help classifiers to learn better and improve generalization, though this effect may vary depending on the specific dataset and context.

Models structure

In this section we will show the structure of two proposed ensemble models, the first one is based on the Bagging classifier concept with a voting classifier, and the second one is based on the stacking classifier concept, the details of both are discussed in the below subsections.

Bagging classifiers with voting classifiers

We present an improved ensemble model that is capable of increasing the model's accuracy and stability by combining different machine-learning techniques through bagging and voting techniques. The ensemble comprises six different classifiers: SVM, DT, Neural Networks (NN), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Naive Bayes (NB). Our model is designed with a bagging framework which, on top of each classifier, improves the generalizability and robustness of the model by applying random subsets of the training data (bootstrapping) and taking the predictions of multiple instances of each classifier.

The bagging idea is about training each base classifier on randomly selected training instances with replacements to form the number of datasets, and training models on these datasets. First, these models make their predictions, then these outputs are averaged to produce a final result by majority voting for classification tasks. This is a method that is mainly used in reducing variance and avoiding the problem of overfitting, which are the biggest problems with complex models such as decision trees and neural networks[51].

Upon the application of bagging classifiers, the ensemble further employs a voting classifier to consolidate the outputs from all base classifiers. This voting classifier operates through two mechanisms: hard voting and soft voting. Hard voting determines the final class prediction by majority rule in Eq. 3. In contrast, soft voting considers the probability estimates p_{ij} for each class j provided by each classifier i, averaging these probabilities to make a prediction in Eq. 4. The structure of both the individual bagging classifiers and the overarching voting mechanism are detailed in Figs. 2 and 3.

$$y_{\text{hard}} = \text{mode}\left\{c_1(x), c_2(x), \ldots, c_N(x)\right\} \quad (3)$$

$$y_{\text{soft}} = \underset{j}{\text{argmax}} \left(\frac{1}{N} \sum_{i=1}^{N} p_{ij}(x) \right) \quad (4)$$

Stacking classifier based on bagging classifiers

Our study goes a step further by implementing a stacking classifier which is an advanced model that exploits the predictive power of multiple classifiers to produce new training data for the final estimator. The stacking model is constructed on four different base classifiers—SVM, NN, DT, and LR, all of which are bagging-based to increase their stability and prevent overfitting. These classifiers are trained on the original training dataset and their predictions (probabilities or classes) serve as new features, which are used to train a second-layer classifier.

The predictions from the bagged versions of SVM, NN, DT, and LR are stacked together to form a new training dataset for the final estimator. In our model, the final estimator is also a Logistic Regression model, applied in a bagging style to maintain consistency with the base classifiers and to further stabilize prediction variance. The structure of the model is detailed in Fig. 4.

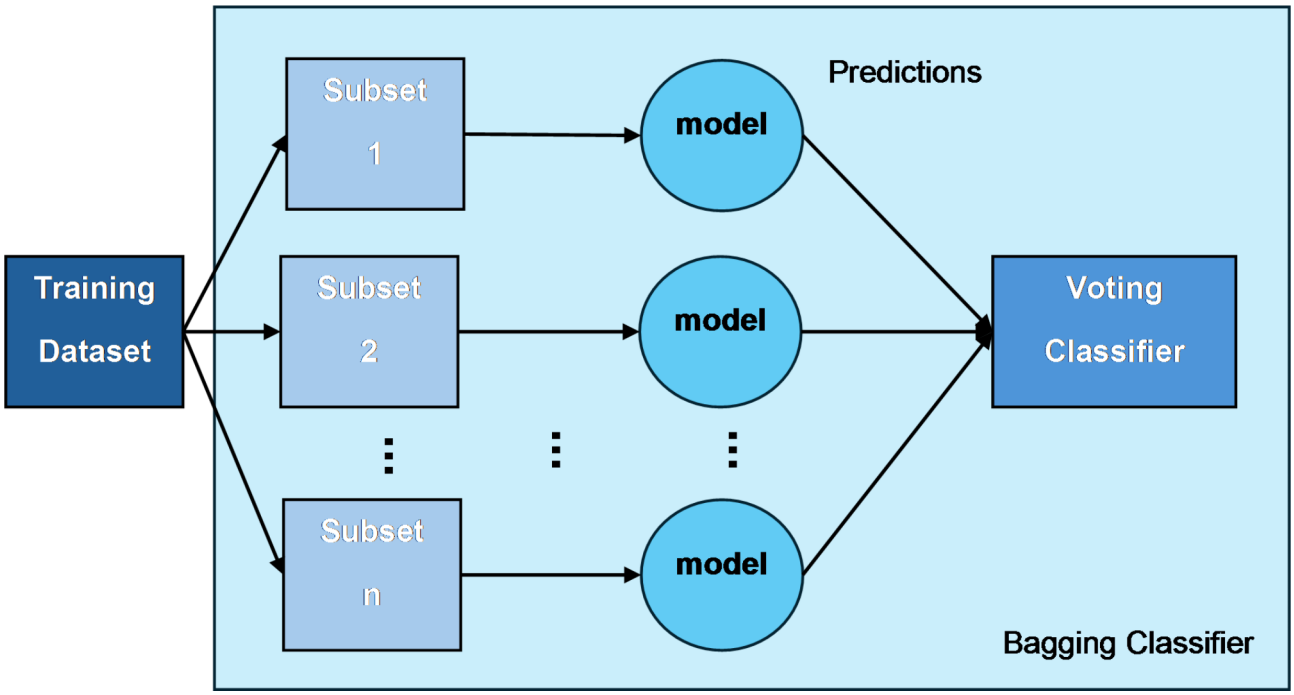

Fig. 2. The structure of the bagging classifier.

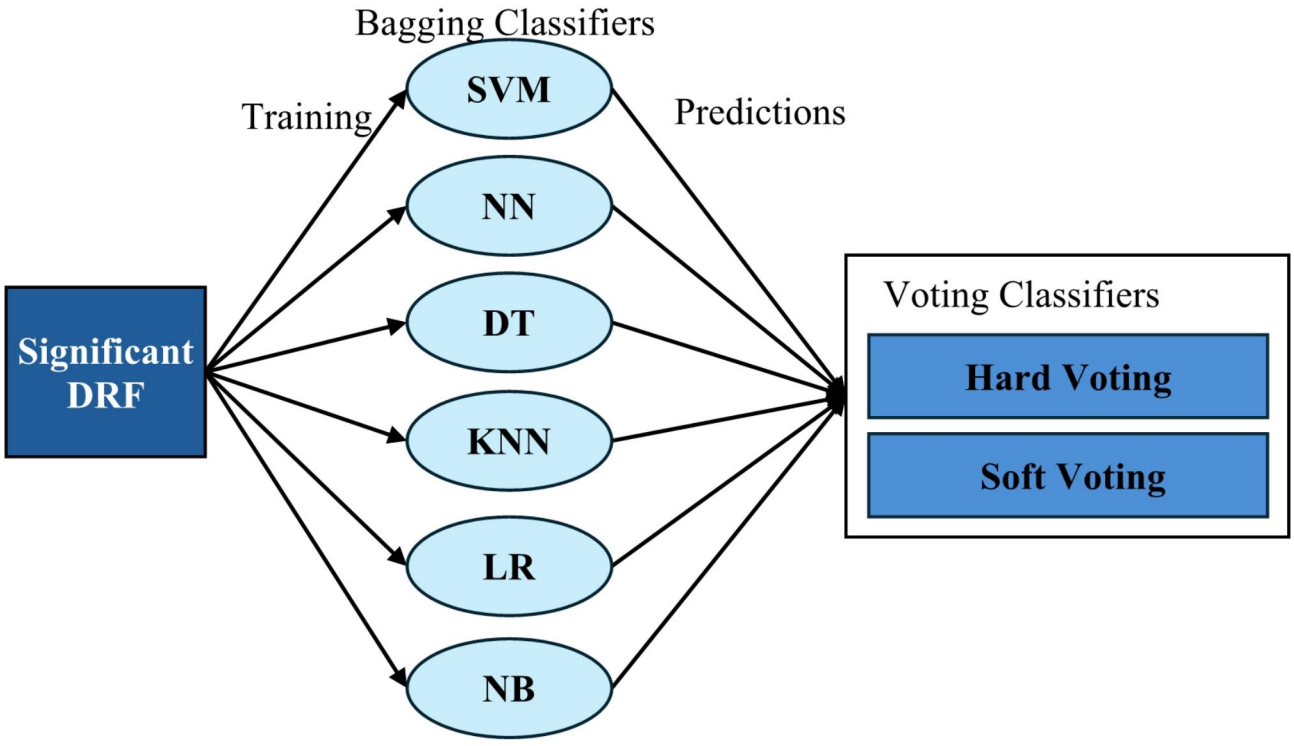

Fig. 3. The structure of the proposed model (bagging classifiers with voting classifiers).

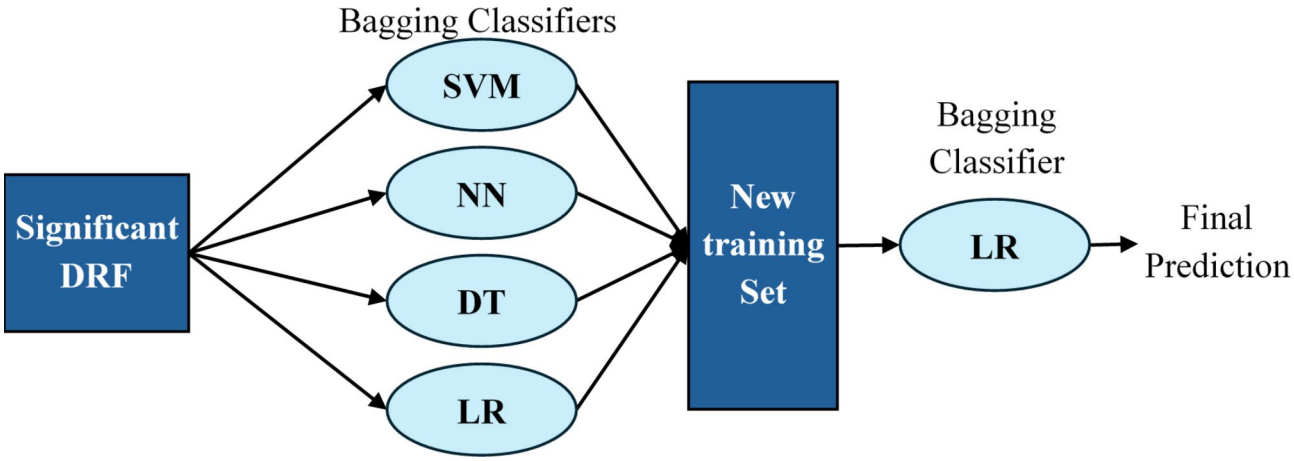

Fig. 4. The structure of the proposed model (stacking classifier based on bagging classifiers (Stacking_ens_LR)).

Performance metrics

To evaluate the performance of our predictive models, we used several metrics including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics were chosen because they provide a comprehensive assessment of the model's performance in terms of both correctness and robustness.

- Accuracy: The ratio of correctly predicted instances to the total instances. It provides a general measure of how well the model performs across all classes.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

- Precision: The ratio of correctly predicted positive observations to the total predicted positives. It indicates the accuracy of the positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

- Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to detect positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

- F1-score: The harmonic mean of precision and recall, providing a single metric that balances both concerns.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

- AUC: AUC measures the ability of the model to distinguish between classes. Higher AUC values indicate better performance.

$$\text{AUC} = \int_0^1 \text{TPR}\,(\text{FPR})\, d\,(\text{FPR}) \tag{9}$$

These metrics were calculated using the scikit-learn library in Python 3.12. The relative importance of these metrics varies depending on the context of the study. In the case of stroke diagnosis, recall (sensitivity) is particularly important as it reflects the model's ability to correctly identify stroke cases, which is crucial for timely and appropriate treatment. Precision is also important to minimize false positives, which can lead to unnecessary interventions. Accuracy provides an overall measure of performance, the F1-score offers a balanced evaluation when precision and recall are equally important, and AUC measures the overall ability of the model to discriminate between positive and negative classes.

Results

The results of this study are systematically divided into three sections: Feature selection, minority oversampling, and the performance of the suggested models. Each section provides a thorough and systematic analysis of the results that were achieved through the methods that were applied.

Feature selection

With the Lasso algorithm, we were able to determine several Dynamic Radiomics Features (DRF) for each evaluation task and indeed boost the performance of the models for this ischemic stroke detection, NIHSS prediction, and outcome prediction tasks. Specifically, 34 DRF were deemed outstanding for ischemic stroke detection, distributed as follows: First_order: 1, GLCM: 9, GLDM: 13, GLRLM: 5, GLSZM: 4, and NGTDM: 2. For the evaluation of NIHSS, 31 DRF were selected, out of which 7 were from First_order, 5 from GLCM, 5 from GLDM, 2 from GLRLM, 4 from GLSZM, and 8 from NGTDM. In the meanwhile, the result prediction task had 40 DRFs chosen with the distribution being 6 in First_order, 21 in GLCM, 9 in GLDM, 2 in GLSZM, 2 in NGTDM, and no DRFs chosen from GLRLM.

Minority over-sampling

To address the class imbalance in the NIHSS prediction and outcome prediction tasks, we have used the SMOTE technique. We employed 80% of the dataset as a training set, and the findings from SMOTE were striking. As shown in Table 3, the number of minority class examples (class 1) increased to 76 and 81 for the NIHSS prediction and outcome prediction tasks, respectively. The number of the majority class (class 0) remained unchanged, and so the balance was achieved without losing the original data integrity.

The performance of the proposed models

The use of feature selection and SMOTE as techniques has considerably improved the efficiency of the proposed two models. The sections below present detailed performance metrics, showing that accuracy, precision, and recall have all increased across all the tasks, thus demonstrating the effectiveness of the integrated approach in dealing with small datasets.

Effect of bagging concept on single model

We made a small comparison to see the effect of the bagging concept on the performance of the single model in different tasks. We calculated the percentage change in performance metrics using Eq. 10. The comparison was made between six models (SVM, DT, NN, KNN, LR, NB) before and after applying bagging. In the stroke detection task, as shown in Fig. 5, the performance of the various models showed both improvements and declines across different metrics. Notably, while most models exhibited improvements in recall, changes in precision varied. The SVM Bagging model shows an increase in recall by 13% and AUC by 4%, but a decrease in precision by 18%. The DT Bagging model demonstrates improvements across all metrics, with a 46% increase in both accuracy and recall. The NN Bagging model shows an increase in recall by 15%, although its precision decreases by 10%. The KNN Bagging model exhibits improvements in AUC and recall by 9% and 20%, respectively, while the LR Bagging model shows a small improvement in accuracy by 4% but a drop in precision by 12%. Lastly, the NB Bagging model shows a rise in recall by 27% and in F1 score by 4%, suggesting enhancements in its predictive reliability after applying bagging. Overall, while the models displayed varied performance changes, the SVM and DT models exhibited the most notable differences.

In the NIHSS prediction task, we found that all models have an improvement in all performance metrics after applying the Bagging, except for the KNN and NB models, which experienced a decrease in precision, as shown in Fig. 6. We also found that recall and F1 score were the most improved metrics due to bagging. The SVM Bagging model shows an increase in AUC by 25% and recall by 59%, along with smaller gains in other metrics. The DT Bagging model exhibits increases across all metrics, with recall and F1 Score increasing by 59% and 53%, respectively. The NN Bagging model shows an increase in AUC and recall by 41%. The KNN Bagging model shows an increase in recall by 39% and a decrease in precision by 19%. The LR Bagging model sees an increase in accuracy by 25% and an increase in F1 Score by 36%. The NB Bagging model sees small gains in most metrics, with AUC and F1 Score increasing by 11% and 9%, respectively.

In the outcome prediction task, applying bagging to different machine learning models results in an increase in performance metrics, as shown in Fig. 7. The NB Bagging model's precision, recall, and F1 score increase by approximately 44%, 14%, and 30%, respectively, while DT Bagging model's AUC, accuracy, precision, recall, and F1 score increase by 38%, 27%, 39%, 88%, 69%. The NN Bagging model shows an increase in accuracy and precision by 11% and 14%, respectively. The KNN Bagging model shows a small increase in accuracy, gaining about 6%. The LR Bagging and SVM Bagging models also show increases: accuracy increases by about 15%, while accuracy and precision of SVM increase by approximately 13% and 40%, respectively.

$$change\,(\%) = \frac{after - before}{before}\% \tag{10}$$

	NIHSS prediction		90-Day mRS	
Label	Before	After	Before	After
1	49	76	44	81
0	76	76	81	81

Table 3. The result of the SMOTE technique.

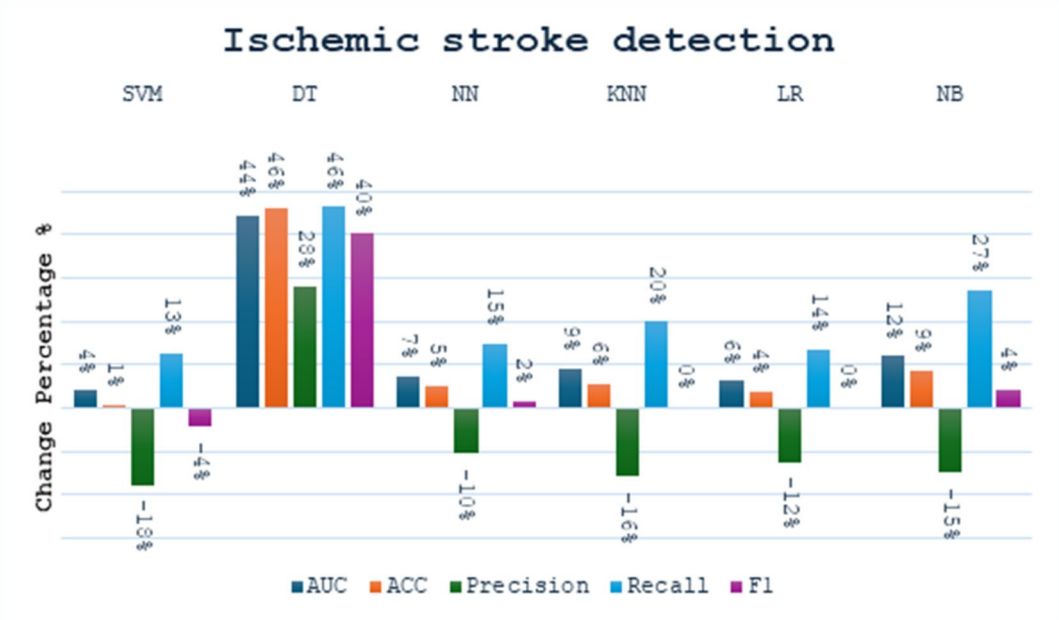

Fig. 5. The change percentage between single models and bagging models at stroke detection task.

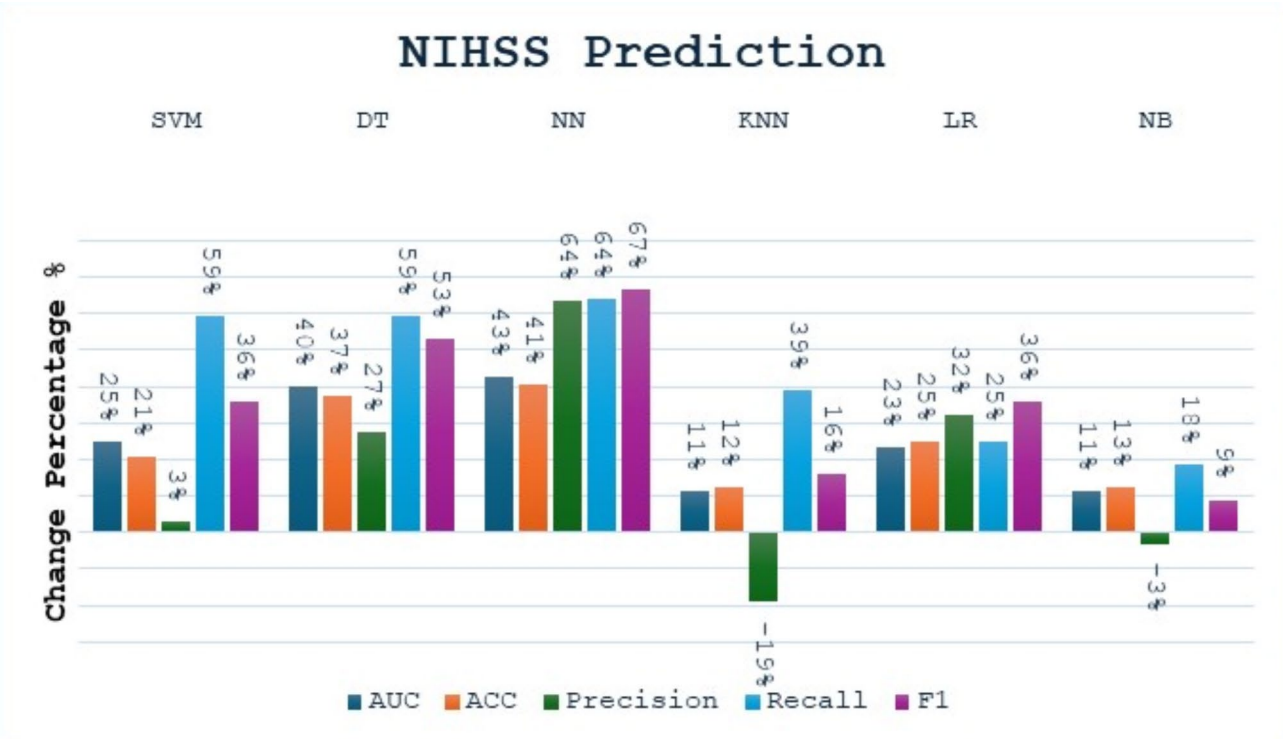

Fig. 6. The change percentage between single models and bagging models at the NIHSS prediction task.

The performance of the proposed models

Table 4 offers a detailed evaluation of various ensemble models applied to three different clinical tasks—ischemic stroke detection, NIHSS prediction, and outcome prediction—highlighting their performance metrics: AUC, accuracy, precision, recall, and F1 Score, without the use of SMOTE technique. Each metric provides insights into the models' abilities to predict accurately and manage errors, with AUC and recall being particularly crucial for understanding model reliability and sensitivity respectively.

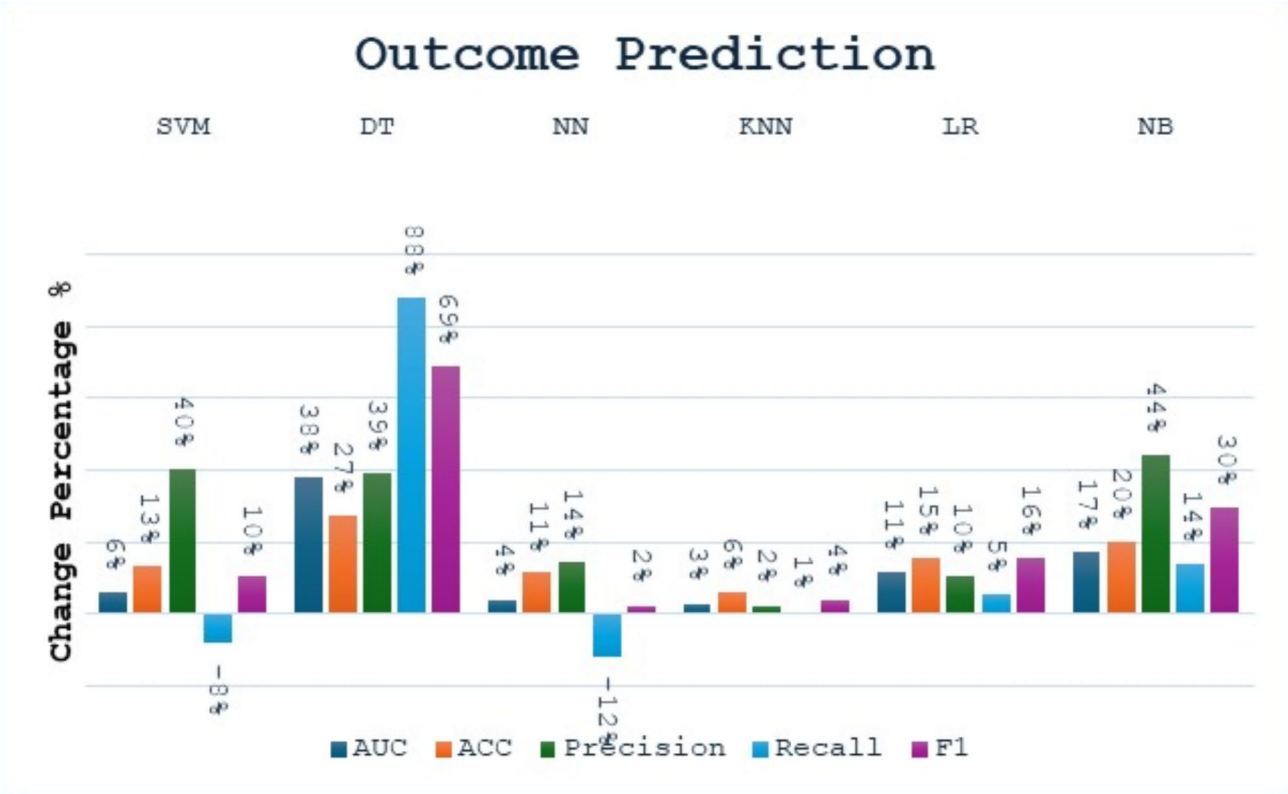

Fig. 7. The change percentage between single models and bagging models at the outcome prediction task.

For ischemic stroke detection, models generally showed robust performance. Most models achieved excellent recall rates of 1.000, indicating perfect sensitivity in identifying true positives—critical for medical diagnoses where missing a true case can be costly. The AUC values were also commendable, with NN Bagging, LR Bagging, Hard Vote, Soft Vote, and Stacking_ens_LR all achieving an AUC of 0.957 or higher, which is considered excellent, reflecting their overall accuracy and reliability in classification across different threshold settings.

In NIHSS prediction, the LR Bagging model stands out with an AUC of 0.938 and a perfect precision of 1.000, though its recall is slightly lower at 0.875. This suggests while the model is excellent at ensuring what it predicts as positive is indeed positive (high precision), it slightly underperforms in capturing all actual positive cases (lower recall). The KNN Bagging model shows weaker performance with the lowest AUC of 0.750 and recall of 0.625, indicating significant room for improvement in both reliability and sensitivity. Stacking_ens_LR performs notably well, achieving the highest AUC of 0.995 among all models evaluated. The model also maintains good accuracy at 0.938 and consistent precision and recall values at 0.875.

For outcome prediction, the variability among models is more pronounced. The highest AUC is noted for the Stacking_ens_LR model at 0.947, suggesting good overall model performance. However, the recall here is lower at 0.556, indicating a weakness in identifying all positive cases, which could be problematic in clinical settings where failing to detect true outcomes can have serious repercussions. Other models display a range of performances. For instance, the SVM Bagging model shows a lower AUC of 0.722 but achieves a high precision of 1.000, indicating it is very accurate when it does predict a positive case, although its lower recall of 0.444 reflects a substantial number of missed true positive cases. On the other hand, NB Bagging has a relatively high AUC of 0.889 and matches its high precision with an F1 Score of 0.875, showcasing better balance in performance metrics compared to some other models.

Table 5 offers a detailed evaluation of various ensemble models applied to three different clinical tasks—ischemic stroke detection, NIHSS prediction, and outcome prediction—highlighting their performance metrics: AUC, accuracy, precision, recall, and F1 Score, after applying the SMOTE technique. In the NIHSS prediction task, the proposed models presented different performance metrics. The SVM Bagging model scored an outstanding AUC of 0.990 while achieving a precision of 0.778 and a recall of 0.875. The best-performing model is LR Bagging with AUC equal to 1.000, precision equal to 1.000, and recall equal to 0.875. The values of such parameters show that the predictive powers are strong and balanced. DT Bagging model is doing great since it obtained an AUC of 0.974, a precision of 0.889, and a perfect recall of 1.000. The NN Bagging model did equally as well by giving 0.995 in AUC, with both precision and recall balanced at 0.875. On the other hand, the KNN Bagging model was good but lower precision. Its tendency was that of false positives. Accordingly, NB Bagging has a slightly low AUC of 0.911, with precision and recall at 0.750, respectively. The Hard Vote and Soft Vote models garnered an AUC of 0.917, with precision at 0.875 and recall at 0.875. The stacking_ens_LR model came out with a high value for AUC of 0.990 and a balanced precision and recall value of 0.875 each.

Task	Model	AUC	Accuracy	Precision	Recall	F1
Ischemic stroke detection	SVM Bagging	0.935	0.906	0.750	1.000	0.857
	DT Bagging	0.901	0.906	0.800	0.889	0.842
	NN Bagging	0.957	0.938	0.818	1.000	0.900
	KNN Bagging	0.935	0.906	0.750	1.000	0.857
	LR Bagging	0.957	0.938	0.818	1.000	0.900
	NB Bagging	0.935	0.906	0.750	1.000	0.857
	Hard Vote	0.957	0.938	0.818	1.000	0.900
	Soft Vote	0.957	0.938	0.818	1.000	0.900
	Stacking_ens_LR	0.966	0.938	0.818	1.000	0.900
NIHSS prediction	SVM Bagging	0.917	0.938	0.875	0.875	0.875
	DT Bagging	0.896	0.906	0.778	0.875	0.824
	NN Bagging	0.917	0.938	0.875	0.875	0.875
	KNN Bagging	0.75	0.813	0.625	0.625	0.625
	LR Bagging	0.938	0.969	1.000	0.875	0.933
	NB Bagging	0.833	0.875	0.75	0.75	0.75
	Hard Vote	0.917	0.938	0.875	0.875	0.875
	Soft Vote	0.917	0.938	0.875	0.875	0.875
	Stacking_ens_LR	0.995	0.938	0.875	0.875	0.875
Outcome Prediction	SVM Bagging	0.722	0.844	1.000	0.444	0.615
	DT Bagging	0.901	0.906	0.800	0.889	0.842
	NN Bagging	0.734	0.813	0.714	0.556	0.625
	KNN Bagging	0.713	0.781	0.625	0.556	0.588
	LR Bagging	0.790	0.844	0.750	0.667	0.706
	NB Bagging	0.889	0.938	1.000	0.778	0.875
	Hard Vote	0.778	0.875	1.000	0.556	0.714
	Soft Vote	0.812	0.875	0.857	0.667	0.750
	Stacking_ens_LR	0.947	0.844	0.833	0.556	0.667

Table 4. The performance metrics of proposed models without applying the SMOTE technique.

Task	Model	AUC	Accuracy	precision	Recall	F1
NIHSS prediction	SVM Bagging	0.990	0.906	0.778	0.875	0.824
	DT Bagging	0.974	0.969	0.889	1.000	0.941
	NN Bagging	0.995	0.938	0.875	0.875	0.875
	KNN Bagging	0.943	0.750	0.500	0.875	0.636
	LR Bagging	1.000	0.969	1.000	0.875	0.933
	NB Bagging	0.911	0.875	0.750	0.750	0.750
	Hard Vote	0.917	0.938	0.875	0.875	0.875
	Soft Vote	0.917	0.938	0.875	0.875	0.875
	Stacking_ens_LR	0.990	0.938	0.875	0.875	0.875
Outcome Prediction	SVM Bagging	0.870	0.781	0.600	0.667	0.632
	DT Bagging	0.899	0.844	0.700	0.778	0.737
	NN Bagging	0.865	0.781	0.583	0.778	0.667
	KNN Bagging	0.894	0.813	0.600	1.000	0.750
	LR Bagging	0.918	0.781	0.583	0.778	0.667
	NB Bagging	0.976	0.906	0.800	0.889	0.842
	Hard Vote	0.824	0.844	0.700	0.778	0.737
	Soft Vote	0.879	0.875	0.727	0.889	0.800
	Stacking_ens_LR	0.903	0.781	0.583	0.778	0.667

Table 5. The performance metrics of proposed models by applying the SMOTE technique.

For the outcome prediction task, the proposed models also demonstrated varied performances. The SVM Bagging model gave an AUC of 0.870, utmost moderate precision, and recall of 0.600, and 0.667, respectively. The bagged DT model gave an AUC of 0.899, with precision being at 0.700 and a balanced performance recall of 0.778. On the other hand, the NN Bagging model gave a lower AUC value of 0.865 but has a moderate precision of 0.583 and a recall of 0.778. The KNN Bagging model had a fairly good AUC of 0.894, with low precision at 0.600 but perfect recall at 1.000. The LR Bagging model presented a high AUC of 0.918, a moderate precision of 0.583, and a recall of 0.778. The highest AUC, strong precision at 0.800, and a recall of 0.889—hence the highest in performance—are what the NB Bagging model has to offer. The Hard Voting model came back with an AUC of 0.824, which produced a balanced precision of 0.700 and a recall of 0.778. On the other side, the Soft Voting model returned quite a good AUC score of 0.879 with a balanced precision of 0.727 and a recall of 0.889. The Stacking_ens_LR model also scored at the top level, with a very strong AUC of 0.903 and balanced precision and recall, 0.583 and 0.778, respectively.

Figure 8 shows the percent variation of the performance metrics for the proposed ensemble models before and after SMOTE is applied to it in the task of NIHSS prediction. SVM Bagging saw an 8% increase in AUC but decreased its accuracy, precision, and F1 score by 3%, 11%, and 6%, respectively. The DT Bagging had improved in all the metrics, notably by 9% in AUC, and a 7% increase in accuracy, among others. Similarly, NN Bagging improved its AUC by 9%. KNN Bagging gave mixed results; while there was an apparent improvement of substantial performance of 26% in AUC and 40% for recall, there was an indication of a notable decrease of 8% in accuracy and 20% for precision. LR Bagging had consistent performance, which showed an improvement of 7% in AUC. However, Hard Vote and Soft Vote models show consistent performance with no changes among any metrics. On the other hand, the Stacking_ens_LR model shows a 1% decrease in AUC, with other metrics being unchanged.

As shown in Fig. 9, the implementation of SMOTE as a preprocessing option for the outcome prediction task resulted in the following percentage changes in performance metrics. The SVM bagging model obtained a 20% increase in AUC, a remarkable increase of 50% in the recall, and a 3% F1 score increase, while accuracy and precision went down by 7% and 40% respectively. DT Bagging model witnessed a reduction across all the metrics, which were the highest of 13% for precision and 12% each for recalls and F1 scores. The NN Bagging model showed an 18% increase in AUC, a 7% increase in F1 score, and a 40% increase in recall, but experienced decreases of 4% in accuracy and 18% in precision. The KNN Bagging model achieved a remarkable 80% increase in recall and a 28% increase in F1 score, besides a 25% rise in AUC, having the drawback of a 4% drop in precision. The NB Bagging model registered improvements of 10%, 14%, %3, %4, and 20% in AUC, accuracy, recall, F1 score, and precision respectively. The Hard Vote model provided a 6% improvement in AUC, a 40% increase in recall, and a 3% increase in F1 score at the expense of precision, which decreased by 30%. The Soft Vote model experienced an 8% increase in AUC, a 33% increase in recall, and a 7% increase in F1 score, while precision dropped by 15%. The Stacking_ens_LR model has shown some loss in AUC, accuracy, and precision, which were 10%, 7%, and 30% respectively, however, recall increased by 40%..

Discussion

The objective of the study is to evaluate the performance of ensemble models, grounded in DRF of DSC-PWI, for ischemic stroke diagnosis, NIHSS prediction, and outcome prediction. The present findings underscore

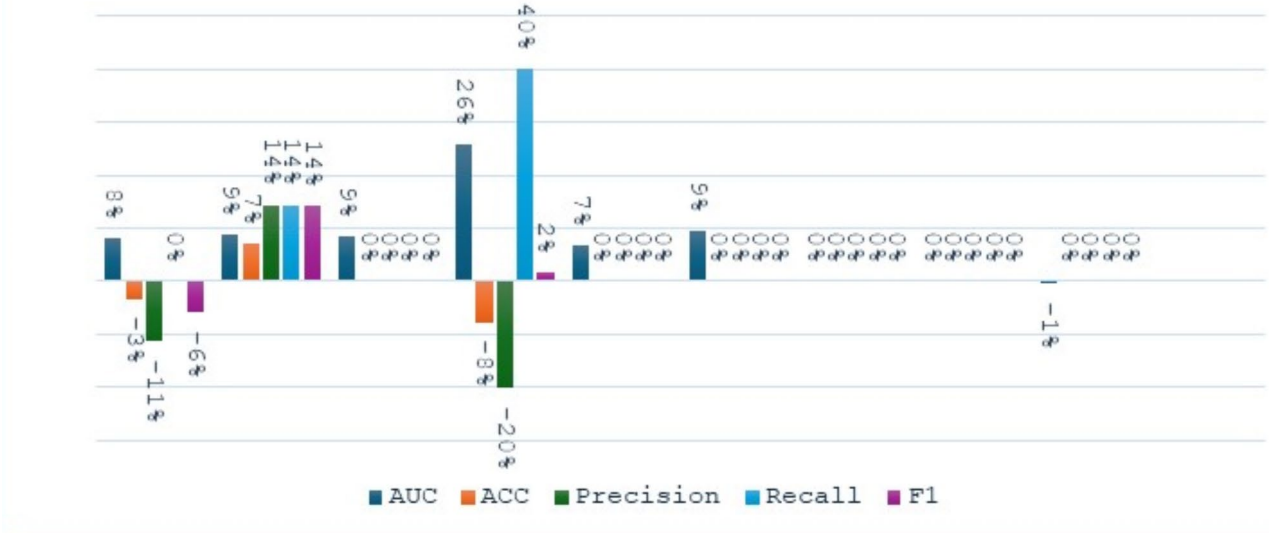

Fig. 8. The change percentage between proposed ensemble models before and after applying SMOTE at NIHSS prediction task.

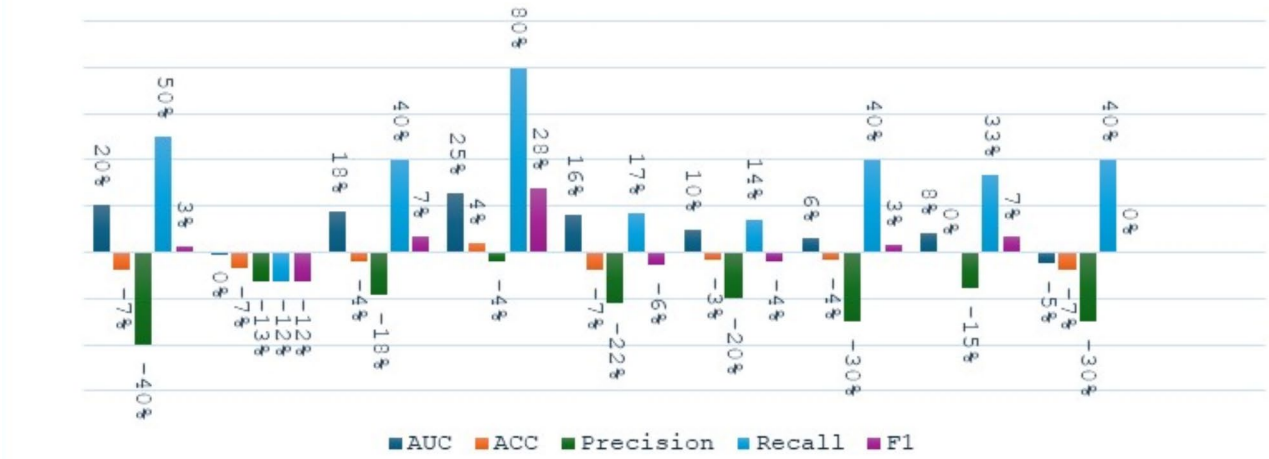

Fig. 9. The change percentage between proposed ensemble models before and after applying SMOTE at the Outcome prediction task.

Task	Model	AUC	precision	Recall
Ischemic stroke detection	NN Bagging	0.957	0.818	1.000
	LR Bagging	0.957	0.818	1.000
	Hard Vote	0.957	0.818	1.000
	Soft Vote	0.957	0.818	1.000
	stacking_ens_LR	0.966	0.818	1.000
	DA(Lasso + PCA + Lasso)[45]	0.925		
NIHSS prediction	NN Bagging	0.995	0.875	0.875
	LR Bagging	1.000	1.000	0.875
	Hard Vote	0.917	0.875	0.875
	Soft Vote	0.917	0.875	0.875
	stacking_ens_LR	0.990	0.875	0.875
	DA(Lasso + PCA + Lasso)[45]	0.853		
Outcome Prediction	NB Bagging	0.976	0.800	0.889
	KNN Bagging	0.894	0.600	1.000
	Soft Vote	0.879	0.727	0.889
	stacking_ens_LR	0.903	0.583	0.778
	LR (Lasso combined)[53]	0.971		
	LR (lasso WB)[53]	0.936		
	CTI + survF [48]	0.949		
	NB(Lasso + PCA + Lasso)[45]	0.828		
	NN(IWS)[52]	0.904		

Table 6. The performance of the proposed model related to related work.

several key aspects concerning the role of ensemble models in the management of ischemic stroke and highlight the utility of specific model structures and preprocessing techniques to better enable accuracy in diagnosis and prognosis. Our results support the ability of the methodology to facilitate predictive analytics in medical diagnostics.

The results presented in Table 6 underscore the significant performance of our proposed ensemble models (best performance bagging classifier, hard vote, soft vote, and Stacking_ens_LR model) across three distinct tasks: ischemic stroke detection, NIHSS prediction, and outcome prediction. In these tasks, the AUC metric was primarily used due to its suitability for binary classification. However, it's important to acknowledge that in clinical settings, sensitivity (recall) and specificity (precision) are often prioritized to minimize false positives and negatives.

For ischemic stroke detection, our ensemble models, including NN Bagging, LR Bagging, Hard Vote, and Soft Vote, achieved a high AUC of 0.957, while the stacking ensemble model (Stacking_ens_LR) slightly outperformed them with an AUC of 0.966. These models also demonstrated high precision (0.818) and perfect recall (1.000),

indicating a strong balance between false positives and negatives. This superior performance, compared to the benchmark model DA (Lasso + PCA + Lasso)[45], which recorded an AUC of 0.925, reaffirms that combining multiple classifiers enhances predictive performance in complex medical imaging tasks. The robustness of these ensemble methods as alternatives for stroke detection is further validated by their consistent high AUC scores.

In the NIHSS prediction task, the LR Bagging model achieved a perfect AUC of 1.000, outperforming the DA (Lasso + PCA + Lasso)[45] benchmark, which managed an AUC of 0.853. The NN Bagging and Stacking_ens_LR models followed closely with AUC scores of 0.995 and 0.990, respectively, showcasing the dominant performance of our models in predicting neurological impairment. The application of the SMOTE technique contributed to a noticeable improvement in AUC, highlighting the importance of addressing class imbalance in predictive models for neurological impairment. These results support the argument that ensemble methods do have substantial improvements in the predictive accuracy of NIHSS. Kindly understand that, based on our findings, we showed that an ensemble model can work with class imbalance; Previous studies have shown that class imbalances can compromise the effectiveness of ensemble methods in various contexts[52,53].

In the outcome prediction task, the NB Bagging model achieved the highest AUC of 0.976, followed by the Stacking_ens_LR model with an AUC of 0.903. The Soft Vote model recorded an AUC of 0.879, and the Hard Vote model had an AUC of 0.824. Compared to related work, such as CTI + survF[48] with an AUC of 0.949 and NN(IWS)[54] with an AUC of 0.904, our proposed models, particularly the NB Bagging, demonstrated superior predictive performance. The LR (Lasso combined)[55] model, which achieved an AUC of 0.971, further validates the robustness of our models. However, the lower recall for some models indicates areas for future improvement, as predicting all positive cases is crucial in clinical situations. These results clearly show the potential benefits of ensemble models, particularly when combined with data preprocessing techniques like SMOTE, in enhancing the prediction performance for patient outcomes. However, the lower recall for some models pinpoints future areas of improvement, as being able to predict all positive cases is crucial in clinical situations.

The application of SMOTE notably improves the AUC and recall across most models, as seen in Table 5. However, this improvement often comes at the cost of reduced precision, indicating an increase in false positives. For NIHSS prediction, after applying SMOTE, models such as SVM Bagging and DT Bagging showed significant improvements in AUC (0.990 and 0.974, respectively) and recall (both 0.875 and 1.000), compared to their performance without SMOTE (AUC of 0.917 and 0.896, recall of 0.875 and 0.875). However, the precision for these models slightly decreased, indicating more false positives. In outcome prediction, the NB Bagging model with SMOTE achieved a high AUC of 0.976 and recall of 0.889, compared to 0.889 and 0.778 without SMOTE. This improvement in recall indicates the model's increased ability to correctly identify positive outcomes, albeit with a slight reduction in precision from 0.750 to 0.800, suggesting an increase in false positives.

The application of the SMOTE technique improves the recall of our models, reducing the number of false negatives. In clinical settings, particularly in long-term outcome predictions (mRS), this can be crucial. A higher recall means fewer missed cases, ensuring that more patients who might benefit from an intervention are correctly identified. This is often more desirable than high precision, which minimizes false positives, because the cost of missing a potential positive case (false negative) can be much higher than incorrectly identifying a non-case as a case (false positive).

The variance observed in Figs. 5, 6 and 7 highlights the differential impact of bagging across various models. These differences are especially notable in the DT model, which exhibits significant performance changes when bagging is applied. Bagging reduces variance by averaging the predictions of multiple trees, thereby stabilizing the model. However, the extent of this stabilization varies across models. For instance, decision trees are inherently high-variance models; thus, bagging them results in substantial performance improvements, as seen in the increased AUC and precision in Figs. 5, 6 and 7. Conversely, models like LR and SVM are less prone to high variance, and bagging has a relatively smaller impact on their performance. This variance indicates that certain models, such as decision trees, benefit more from bagging, which justifies their inclusion in ensemble methods for our predictive model. In contrast, methods that show minimal improvement or even performance degradation with bagging, such as KNN, may be less suitable for this approach. This observation can guide the selection of models for ensemble methods, favoring those with higher variability reduction through bagging.

The findings of this research have serious clinical implications. Among ensemble models, stacking and bagging approaches have shown great potential in predictive accuracy for ischemic stroke diagnosis, NIHSS prediction, and further outcome prediction. These could be potential models to improve patient management since they offer the possibility to make assessments more accurate and individualized. In the case of precision medicine in stroke care, a combined model of DRF and Ensemble models has good potential for application.

Our ensemble-based approach showed robust performance in ischemic stroke diagnosis, NIHSS prediction, and outcome prediction. However, there are several limitations to this approach. First, the size of the sample used was relatively small, though it was sufficient for the first assessment. Generally, this relatively small sample size would have proved difficult to make proper generalizations about the findings and, still, capture the diverse patient populations in which these tests are used in clinical practice. Moreover, DSC-PWI being valuable by itself may not necessarily represent the whole information content that other imaging modalities or clinical data might provide for diagnosis and prognosis.

This would, in the future, be able to validate in a bigger, prospective cohort to have an assessment of generalizability and robustness of the ensemble models. This sometimes also may require multicenter studies to factor in the differences between patient demographics and diversity in imaging techniques. Moreover, future work may include multimodal imaging and clinical data as features, which would likely enhance the predictive performance. Lastly, advanced interpretability techniques for ensemble models should be explored, since understanding the decision process of ensembles is deemed crucial for clinical adoption.

Conclusion

In this study, we established the validity of the DRF ensemble method as a diagnostic tool for ischemic stroke, for NIHSS prediction, and for predicting outcomes. Our presented models, particularly the stacking_ens_LR and NB Bagging models, scored high AUCs, which were higher or at least equal to the benchmark models proposed in the related works. The results give a clear example of the role of ensemble learning in improving diagnostic and prognostic precision in ischemic stroke. Utilizing feature selection, SMOTE, and creative ensemble methods we were able to properly address issues like class imbalance and high dimensions and demonstrated the effectiveness of those methods for advancing precision medicine in stroke treatment. Despite some restraints, the conclusions that were made here demonstrate the part that ensemble models can play in stroke management and are on the way to even better outcomes. These ensemble models can be used instead of Single models in Radiomics study.

Data availability

All data sets generated during and/or analyzed during the present study are not publicly available but processed dataset are available from the corresponding author based on reasonable scientific merit. All data provided are anonymized to respect the privacy of the participants who participated in the study.

Received: 16 May 2024; Accepted: 30 October 2024
Published online: 11 November 2024

References

1. Fan, J. et al. Global Burden, Risk Factor Analysis, and Prediction Study of Ischemic Stroke, 1990–2030, *Neurology*, vol. 101, no. 2, doi: (2023). https://doi.org/10.1212/WNL.0000000000207387
2. Nam, H. S. & Kim, B. M. Advance of Thrombolysis and Thrombectomy in Acute ischemic stroke. *J. Clin. Med.* **12** (2). https://doi.org/10.3390/jcm12020720 (2023).
3. Cynthia & Aulia, D. Ischemic stroke with anticoagulant protein C Deficiency. *Int. J. Sci. Soc.* **5** (1). https://doi.org/10.54783/ijsoc.v5i1.657 (2023).
4. Kong, J., Chu, R. & Wang, Y. Neuroprotective treatments for ischemic stroke: opportunities for Nanotechnology. *Adv. Funct. Mater.* **32** (52). https://doi.org/10.1002/adfm.202209405 (2022).
5. Clary, B. L. et al. Abstract 19: Loss Of Endothelial Tissue-nonspecific Alkaline Phosphatase Modifies Sensorimotor Deficits In Chronic Ischemic Stroke, *Stroke*, vol. 54, no. Suppl_1, doi: (2023). https://doi.org/10.1161/str.54.suppl_1.19
6. Williams, D. M. & Felix, A. C. G. Prevention, diagnosis, and management of stroke. in *Reichel's Care Elder.*, (2022).
7. Anton-Munarriz, C. et al. Detection of cerebral ischaemia using transfer learning techniques, in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2023-June, doi: (2023). https://doi.org/10.1109/CBMS58004.2023.00284
8. Guo, X. & Dye, J. Modern Prehospital Screening Technology for Emergent Neurovascular disorders. *Adv. Biology*. **7** (10). https://doi.org/10.1002/adbi.202300174 (2023).
9. Alshehri, F. Imaging based detection of Acute Ischemic Stroke Via Multidetector Computed Tomography. *J. Umm Al-Qura Univ. Med. Sci.* **9** (1). https://doi.org/10.54940/ms94397891 (2023).
10. Ma, X. et al. Evaluation of infarct core and ischemic penumbra by absolute quantitative cerebral dynamic susceptibility contrast perfusion magnetic resonance imaging using self-calibrated echo planar imaging sequencing in patients with acute ischemic stroke. *Quant. Imaging Med. Surg.* **12** (8). https://doi.org/10.21037/qims-21-975 (2022).
11. Niibo, T. et al. Arterial spin-labeled perfusion imaging to predict mismatch in acute ischemic stroke. *Stroke*. **44** (9). https://doi.org/10.1161/STROKEAHA.113.002097 (2013).
12. Amukotuwa, S. A. et al. Comparison of T2*GRE and DSC-PWI for hemorrhage detection in acute ischemic stroke patients: pooled analysis of the EPITHET, DEFUSE 2, and SENSE 3 stroke studies. *Int. J. Stroke*. **15** (2). https://doi.org/10.1177/1747493019858781 (2020).
13. Liu, J., Lin, C., Minuti, A. & Lipton, M. Arterial spin labeling compared to dynamic susceptibility contrast MR perfusion imaging for assessment of ischemic penumbra: a systematic review. *J. Neuroimaging*. **31** (6). https://doi.org/10.1111/jon.12913 (2021).
14. Yao, G., Zhang, J., Yu, M., Yang, Z. & Chu, H. Factors affecting the prognosis of patients with Acute Cerebrovascular occlusion with High National Institutes of Health Stroke Scale scores treated with SWIM Technology. *Altern. Ther. Health Med.*, **29**, 6, (2023).
15. Kwah, L. K. & Diong, J. National Institutes of Health Stroke Scale (NIHSS). *J. Physiotherapy*. **60** (1). https://doi.org/10.1016/j.jphys.2013.12.012 (2014).
16. Pratama, L. B. et al. IN A TERTIARY HOSPITAL. *MNJ (Malang Neurol. Journal)*. **9** (1). https://doi.org/10.21776/ub.mnj.2023.009.01.4 (2022).
17. Yoo, A. J. et al. Combining acute diffusion-weighted imaging and mean transmit time lesion volumes with national institutes of health stroke scale score improves the prediction of acute stroke outcome. *Stroke*. **41** (8). https://doi.org/10.1161/STROKEAHA.110.582874 (2010).
18. Campagnini, S. et al. Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review. *J. Neuroeng. Rehabil.* **19** (1). https://doi.org/10.1186/s12984-022-01032-4 (2022).
19. Fast, L. et al. Machine learning-based prediction of clinical outcomes after first-ever ischemic stroke. *Front. Neurol.* **14** https://doi.org/10.3389/fneur.2023.1114360 (2023).
20. Stinear, C. M., Smith, M. C. & Byblow, W. D. Prediction tools for Stroke Rehabilitation. *Stroke*. **50** (11). https://doi.org/10.1161/STROKEAHA.119.025696 (2019).
21. Wu, O. et al. Role of Acute Lesion Topography in initial ischemic stroke severity and long-term functional outcomes. *Stroke*. **46** (9). https://doi.org/10.1161/STROKEAHA.115.009643 (2015).
22. Douiri, A. et al. Patient-specific prediction of functional recovery after stroke. *Int. J. Stroke*. **12** (5). https://doi.org/10.1177/1747493017706241 (2017).
23. Yan, C. et al. Development and validation of a nomogram model for predicting unfavorable functional outcomes in ischemic stroke patients after acute phase. *Front. Aging Neurosci.* **15** https://doi.org/10.3389/fnagi.2023.1161016 (2023).
24. Cramer, S. C. et al. Intense Arm Rehabilitation Therapy improves the Modified Rankin Scale score: Association between gains in impairment and function. *Neurology*. **96** (14). https://doi.org/10.1212/WNL.0000000000011667 (2021).
25. Campana, A., Gandomkar, Z., Giannotti, N. & Reed, W. The use of radiomics in magnetic resonance imaging for the pre-treatment characterisation of breast cancers: a scoping review. *J. Med. Radiat. Sci.* **70** (4). https://doi.org/10.1002/jmrs.709 (2023).
26. Kang, W. et al. Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis. *J. Translational Med.* **21** (1). https://doi.org/10.1186/s12967-023-04437-4 (2023).
27. Scapicchio, C. et al. A deep look into radiomics. *Radiologia Med.* **126** (10). https://doi.org/10.1007/s11547-021-01389-x (2021).
28. Polidori, T. et al. Radiomics applications in cardiac imaging: a comprehensive review. *Radiol. Med.* **128** (8). https://doi.org/10.1007/s11547-023-01658-x (2023).

29. Wu, H. et al. Radiomics analysis of the optic nerve for detecting dysthyroid optic neuropathy, based on water-fat imaging. *Insights Imaging*. **13** (1). https://doi.org/10.1186/s13244-022-01292-7 (2022).
30. Carrera-Escalé, L. et al. Radiomics-Based Assessment of OCT angiography images for Diabetic Retinopathy diagnosis. *Ophthalmol. Sci.* **3** (2). https://doi.org/10.1016/j.xops.2022.100259 (2023).
31. Guo, J. et al. MR-based radiomics signature in differentiating ocular adnexal lymphoma from idiopathic orbital inflammation. *Eur. Radiol.* **28** (9). https://doi.org/10.1007/s00330-018-5381-7 (2018).
32. Li, Z., Guo, J., Xu, X., Wei, W. & Xian, J. MRI-based radiomics model can improve the predictive performance of postlaminar optic nerve invasion in retinoblastoma. *Br. J. Radiol.* **95** (1130). https://doi.org/10.1259/bjr.20211027 (2022).
33. Russo, L., Charles-Davies, D., Bottazzi, S., Sala, E. & Boldrini, L. Radiomics for clinical decision support in radiation oncology. *Clin. Oncol.* **36** (8). https://doi.org/10.1016/j.clon.2024.03.003 (2024).
34. Dragoș, H. M. et al. MRI Radiomics and Predictive models in assessing ischemic stroke Outcome—A systematic review. *Diagnostics*. **13** (5). https://doi.org/10.3390/diagnostics13050857 (2023).
35. Wen, X., Hu, X., Xiao, Y. & Chen, J. Radiomics analysis for predicting malignant cerebral edema in patients undergoing endovascular treatment for acute ischemic stroke. *Diagn. Interv Radiol.* **29** (2). https://doi.org/10.4274/dir.2023.221764 (2023).
36. Singh, U., Jena, A. K. & Haque, M. T. An Ensemble Learning Approach and Analysis for Stroke Prediction Dataset, doi: (2022). https://doi.org/10.1109/ASSIC55218.2022.10088363
37. Alruily, M., El-Ghany, S. A., Mostafa, A. M., Ezz, M. & El-Aziz, A. A. A. A-Tuning ensemble machine learning technique for cerebral stroke prediction. *Appl. Sci.* **13** (8). https://doi.org/10.3390/app13085047 (2023).
38. Ye, W. et al. OEDL: an optimized ensemble deep learning method for the prediction of acute ischemic stroke prognoses using union features. *Front. Neurol.* **14** https://doi.org/10.3389/fneur.2023.1158555 (2023).
39. Gottam, B., Mandula, L., Kanaparthi, A., Kumar, D. K. K. & Chavan, G. B. Ensemble-based AI system for Brain Stroke Prediction. *Int. J. Res. Appl. Sci. Eng. Technol.* **11** (6). https://doi.org/10.22214/ijraset.2023.53345 (2023).
40. Lee, S. et al. May,., Ensemble learning-based radiomics with multi-sequence magnetic resonance imaging for benign and malignant soft tissue tumor differentiation, *PLoS One*, vol. 18, no. 5 doi: (2023). https://doi.org/10.1371/journal.pone.0286417
41. Yu, H. et al. Prognosis of ischemic stroke predicted by machine learning based on multi-modal MRI radiomics. *Front. Psychiatry*. **13** https://doi.org/10.3389/fpsyt.2022.1105496 (2023).
42. Gerbasi, A. et al. Prognostic value of combined Radiomic features from Follow-Up DWI and T2-FLAIR in Acute ischemic stroke. *J. Cardiovasc. Dev. Dis.* **9** (12). https://doi.org/10.3390/jcdd9120468 (2022).
43. Liu, J. et al. Prediction of recurrence of ischemic stroke within 1 year of discharge based on machine learning MRI radiomics. *Front. Neurosci.* **17** https://doi.org/10.3389/fnins.2023.1110579 (2023).
44. Shree, R. et al. Application of Ensemble Methods in Medical Diagnosis,., (2023).
45. Guo, Y. et al. A focus on the role of DSC-PWI dynamic Radiomics features in diagnosis and outcome prediction of ischemic stroke. *J. Clin. Med.* **11** (18). https://doi.org/10.3390/jcm11185364 (2022).
46. Smith, S. M. et al. Advances in functional and structural MR image analysis and implementation as FSL, in *NeuroImage*, vol. 23, no. SUPPL. 1, doi: (2004). https://doi.org/10.1016/j.neuroimage.2004.07.051
47. Fan, S. et al. An automatic estimation of arterial input function based on multi-stream 3d CNN. *Front. Neuroinform.* **13** https://doi.org/10.3389/fninf.2019.00049 (2019).
48. Guo, Y. et al. Novel survival features generated by clinical text information and Radiomics features may improve the prediction of ischemic stroke outcome. *Diagnostics*. **12** (7). https://doi.org/10.3390/diagnostics12071664 (2022).
49. Cai, T. Breast Cancer diagnosis using Imbalanced Learning and Ensemble Method. *Appl. Comput. Math.* **7** (3). https://doi.org/10.11648/j.acm.20180703.20 (2018).
50. Elreedy, D. & Atiya, A. F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf. Sci. (Ny).* **505** https://doi.org/10.1016/j.ins.2019.07.070 (2019).
51. Ponnaganti, N. D. & Anitha, R. A Novel Ensemble Bagging Classification Method for Breast Cancer Classification Using Machine Learning Techniques, *Trait. du Signal*, vol. 39, no. 1, pp. 229–237, Feb. doi: (2022). https://doi.org/10.18280/ts.390123
52. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man. Cybernetics Part. C: Appl. Reviews*. **42** (4). https://doi.org/10.1109/TSMCC.2011.2161285 (2012).
53. Li, J., Du, J. & Zhang, X. A Clustering Resampling Stacked Ensemble Method for Imbalance Classification Problem, in *IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, 2022, pp. 741–748, doi: (2022). https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00124
54. Lu, J. et al. Determining acute ischemic stroke onset time using machine learning and radiomics features of infarct lesions and whole brain. *Math. Biosci. Eng.* **21** (1), 34–48. https://doi.org/10.3934/mbe.2024002 (2023).
55. Guo, Y. et al. The combination of whole-brain features and local-lesion features in DSC-PWI May improve ischemic stroke outcome prediction. *Life*. **12** (11). https://doi.org/10.3390/life12111847 (2022).

Author contributions

M.M.Y., J.L., and A.Z. designed the study and conceptualized the research. H.Y. and A.C. performed the experiments and collected the data. M.M.Y., X.Z. and H.H. analyzed the data. T.H., X.M., and Y.S. contributed to the interpretation of the results. Y.G. and Y.L. prepared the figures. M.M.Y., J.L., and Y.K. wrote the main manuscript text. All authors reviewed and approved the final manuscript.

Funding

This research was funded by the National Key Research and Development Program of China, grant number 2022YFF0710800; the National Key Research and Development Program of China, grant number 2022YFF0710802; the National Natural Science Foundation of China, grant number 62071311.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78353-y.

Correspondence and requests for materials should be addressed to Y.K.

Competing interests.
The authors declare no competing interests.