scientific reports

OPEN

Predictive modelling and identification of key risk factors for stroke using machine learning

Ahmad Hassan[1], Saima Gulzar Ahmad[1], Ehsan Ullah Munir[1], Imtiaz Ali Khan[2] & Naeem Ramzan[3]✉

Strokes are a leading global cause of mortality, underscoring the need for early detection and prevention strategies. However, addressing hidden risk factors and achieving accurate prediction become particularly challenging in the presence of imbalanced and missing data. This study encompasses three imputation techniques to deal with missing data. To tackle data imbalance, it employs the synthetic minority oversampling technique (SMOTE). The study initiates with a baseline model and subsequently employs an extensive range of advanced models. This study thoroughly evaluates the performance of these models by employing k-fold cross-validation on various imbalanced and balanced datasets. The findings reveal that age, body mass index (BMI), average glucose level, heart disease, hypertension, and marital status are the most influential features in predicting strokes. Furthermore, a Dense Stacking Ensemble (DSE) model is built upon previous advanced models after fine-tuning, with the best-performing model as a meta-classifier. The DSE model demonstrated over 96% accuracy across diverse datasets, with an AUC score of 83.94% on imbalanced imputed dataset and 98.92% on balanced one. This research underscores the remarkable performance of the DSE model, compared to the previous research on the same dataset. It highlights the model's potential for early stroke detection to improve patient outcomes.

Stroke, a devastating medical condition, is a leading cause of mortality worldwide. It occurs when the blood supply to the brain is interrupted or reduced, impairing brain functions[1]. As per the World Stroke Organization (WSO), there is a significant risk associated with strokes, with one in four individuals over the age of 25 facing the possibility of experiencing a stroke during their lifetime[2]. Stroke is a common condition that significantly affects the population. Stroke is the second most common cause of death and the third most prevalent reason for impairment in adults globally. It is a major factor in both death and disability[3]. The significant impact of chronic illness on people, families, and healthcare systems highlights the need for precise and timely prediction techniques to enhance patient outcomes[4].

In the field of medicine, machine learning has become a powerful technology that has the potential to transform stroke prevention and prediction[5–7]. Machine learning models use large datasets and sophisticated algorithms to identify hidden risk factors, forecast outcomes, and offer tailored strategies for treatment[8]. Stroke prediction is a vital area of research in the medical field. However, there are several problems and issues that need to be resolved[9,10]. The accuracy of predictive models is one of the main issues. Machine learning models have shown potential in stroke prediction. Factors such as the data quality, the choice of features, and the choice of algorithm can impact how well models perform[11]. To ensure these model's dependability and efficacy in predicting strokes, it is crucial to assess and validate these factors[12] carefully. Another critical concern is the handling of missing data. Predictive prediction model performance can be severely impacted by incomplete data, producing erroneous or biased outcomes[13]. Appropriate data imputation approaches are needed to handle missing data and increase the precision of prediction models[14].

The data imbalance is also a concern in stroke prediction[15]. Due to the rarity of pre-stroke datasets, they frequently contain imbalanced classifications, with most instances being non-stroke cases[16]. This imbalance can result in biased models that favour the majority and ignore the minority, resulting in low forecast accuracy. To solve this issue and increase the effectiveness of prediction models, several oversampling and undersampling methods are employed, the popular of which is the SMOTE[17,18]. Furthermore, due to ethical considerations, it

[1]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Grand Trunk Road, Wah 47010, Pakistan. [2]Department of Computer Science, Cardiff School of Technologies, Llandaff Campus, Western Avenue, Cardiff CF5 2YB, UK. [3]School of Computing, Engineering and Physical Sciences, University of the West of Scotland, High Street, Paisley PA1 2BE, UK. ✉email: naeem.ramzan@uws.ac.uk

is challenging to get stroke prediction datasets, especially regarding patient privacy[19]. Predictive models that employ sensitive health data must follow strong privacy standards and protect patients' rights and autonomy[20–22].

Overall, stroke prediction is a complex and challenging area of study that demands careful evaluation of numerous challenges and concerns[23]. However, by developing innovative approaches and employing rigorous evaluation methods, the potential of machine learning in stroke prediction can be fully realized[24]. These approaches and methods can improve patient outcomes and lower the societal and individual burden of stroke[25]. Addressing stroke prediction difficulties such as accuracy, missing data, data imbalance, and interpretability is critical to reaching the full potential of machine learning in this domain[26].

The alarming statistics and various issues highlight the urgent need for effective stroke prevention and prediction strategies. This research endeavour delves into the realm of advanced machine learning models to predict strokes and identify key risk factors. By harnessing the power of these models, it aims to enhance early detection, minimize the impact of strokes, and ultimately improve patient outcomes. This study uses a comprehensive analysis of various machine-learning models to predict strokes. It makes several contributions to stroke prediction and provides various previously unknown insights, including:

- Exploring various data imputation techniques and addressing data imbalance issues in order to enhance the accuracy and robustness of stroke prediction models.
- Identifying crucial features for stroke prediction and uncovering previously unknown risk factors, giving a comprehensive understanding of stroke risk assessment.
- Creating an augmented dataset incorporating important key risk factor features using the imputed datasets, enhancing the effectiveness of stroke prediction models.
- Assessing the effectiveness of advanced machine learning models across different datasets and creating a robust Dense Stacking Ensemble model for stroke prediction.
- The key contribution is showcasing the enhanced predictive capabilities of the model in accurately identifying and testing strokes, surpassing the performance of prior studies that utilized the same dataset.

These contributions collectively enhance the overall understanding of stroke prediction and key contributing factors for stroke. It highlights the potential of machine learning models in accurately identifying individuals at risk of strokes. The literature review can be found in Section "Literature review". Moving on to Section "Dataset and preprocessing", which delves into examining the dataset used, the challenges that arise while data preparation, and the preprocessing strategies employed. Section "Data modelling" provides an overview of the main research workflow and outlines the approach to its execution. It examines many models for data modelling. Machine learning algorithms that are used for forecasting are discussed in Section "Machine learning algorithms". Section "Results" presents an overview of the prediction results obtained through the utilization of different machine learning models and approaches, along with a discussion subsection. Meanwhile, Section "Conclusion" encompasses a conclusion of the findings along with recommendations for future research endeavours.

Literature review

The field of stroke prediction research has been the subject of numerous contributions by various authors over an extended period that uses various datasets. However, in this paper, recent contributions are focused that utilize the same dataset as these are also used for evaluation as well. Several machine learning models, including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression, are used to predict stroke. The authors also propose a Minimal Genetic Folding (MGF) model[27] for predicting the probability of stroke, achieving an accuracy of 83.2%. The MGF classification is the most accurate, surpassing the area under the curve (AUC) scores of the other specified kernels. The research supports the notion that a general MGF kernel could differentiate between various stages of stroke recovery, but more research is needed. The study's potential limitations include the oversampling method, which might have affected how well the MGF classifier performed.

The authors propose a strategy for predicting stroke using a Logistic Regression algorithm. The authors employ preprocessing techniques such as SMOTE, feature selection, and controlling outliers to enhance the model's performance[28]. By analyzing various factors such as blood pressure, body mass, heart conditions, age, previous smoking status, prior history of stroke, and glucose levels, the authors achieve an accuracy of 86% in stroke disease prediction, which outperforms other LR-based models. The research emphasizes the capacity of machine learning methods to reduce the adverse impacts of stroke and enable early detection. Multiple physiological attributes are used in various machine learning techniques to forecast strokes, that includes Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours, Support Vector Machine, and Naive Bayes based[29]. The findings indicate that Naïve Bayes achieves the highest accuracy rate, reaching around 82%. These findings suggest that machine learning models can aid early stroke identification in the future.

To predict strokes and evaluate, the proposed model achieves 94% accuracy, and the model outperforms other algorithms, including Naive Bayes, Logistic Regression, Support Vector Machine, and Decision Tree[30]. The authors also used Ensembled Naive Bayes and Ensembled Decision Tree. Overall, the article's contributions to developing an integrated learning model and reorganizing the fixed structure of the developed algorithm.

The researchers employ machine learning algorithms for predicting stroke and evaluate their performance based on F1 score, recall, accuracy, and precision[31]. Preprocessing steps include handling missing values, one hot encoding, and feature scaling. The authors use three classifiers, Support Vector Machines, Decision Trees, and Logistic Regression, to train on the dataset and compare their results. The study emphasizes the value of early stroke prediction, and the paper's contribution lies in preparing the dataset using machine learning algorithms. The proposed model achieves an accuracy of 95.49% and can be used for early stroke prediction in real-world applications.

In another study, the authors put forth a predictive model for stroke detection using five different algorithms, i.e. K-Nearest Neighbours, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression[32]. A comparative analysis of the five models reveals that Random Forest has the highest accuracy of 95.5%. The authors conclude that they find that Random Forest is the model with the highest accuracy and fewest false negatives, and they use Tkinter to construct a Graphical User Interface (GUI) to make the application's use more convenient. The authors suggest that more medical attributes should be considered in future work for better performance of the model. Four distinct models are utilized, including Logistic Regression, Voting classifier, Decision Tree, and Random Forest[33]. Random Forest performs better by achieving the highest classification accuracy of 96%. The future scope of their research involves using more extensive datasets and different machine learning methods, such as AdaBoost, SVM, and bagging, to enhance prediction reliability further. The authors suggest that machine learning can aid patients in receiving early stroke treatment and enhance their quality of life.

In another article[34], the authors explore the performance of Logistic Regression and Random Forest algorithms in predicting strokes using a preprocessed stroke dataset. The Random Forest algorithm outperforms Logistic Regression in terms of accuracy. The study also discusses the bias and variance of the models and their impact on the results. Although there are some limitations in the proposed work, such as only using two models are used however it provides valuable insight into stroke prediction research. A recent study suggests an ensemble RXLM model to predict stroke using Random Forest, XGBoost, and LightGBM[35]. The dataset is pre-processed using the KNN imputer technique, one-hot encoding, and SMOTE. The researchers fine-tune the hyperparameters of the ML algorithms by employing a random search technique to achieve optimal parameter values. The accuracy of the suggested ensemble RXLM model is 96.34%.

Authors in their study propose a machine learning model with K-Nearest Neighbours, Decision Tree, and Logistic Regression[36]. Exploratory data analysis is applied for preprocessing and uses the SMOTE technique to balance the dataset. Finally, a cloud-based mobile app is developed, which can gather user data for analysis and accurately warn a person of the likelihood of a stroke with an accuracy of 96%. Future work will focus on analyzing the dataset using deep learning methods to enhance accuracy. The authors explore ten machine-learning models to predict strokes[37]. The employed models included Gaussian Naive Bayes, Bernoulli Naive Bayes, Gradient Boosting, Stochastic Gradient Descent, K-nearest neighbours (KNN), support vector machine (SVM), Decision Tree, Random Forest, Logistic Regression, and MLP (Multi-Layer Perceptron). The study emphasizes the necessity of an early stroke diagnosis to lessen its effects with a 94% accuracy rate, the KNN algorithm outperforms the other models.

The literature review explores various machine learning models for stroke prediction that include Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression. The studies also propose new models, highlighting the importance of early detection and achieving accuracy rates ranging from 82 to 96%. However, limitations such as limited model selection, feature selection/ engineering, and dataset size are identified. This research paper addresses these deficiencies by conducting a comprehensive analysis of advanced machine learning models, identifying key risk factors along with importance, and evaluating the performance on a larger augmented dataset.

Dataset and preprocessing

The stroke prediction dataset was created by McKinsey & Company and Kaggle is the source of the data used in this study[38,39]. The dataset is in comma separated values (CSV) format, including demographic and health-related information about individuals and whether or not they have had a stroke. The dataset was originally comprised a total of 29,072 records, while only 30% of the data is publicly accessible and the remaining 70% is designated as private data[40]. The source of the dataset is mentioned as confidential. The data originates from medical records associated with 5110 individuals residing in Bangladesh. The dataset has underwent preprocessing procedures, which involved modifications to the original dataset sourced from Electronic Health Records (EHR) managed by McKinsey & Company[41]. The data has some missing values, and there is an imbalance between the number of people who have had a stroke and those who have not. The aim is to address these issues using different data imputation techniques and oversampling methods.

Exploratory analysis

The dataset used for stroke prediction consists of 5110 observations, each containing 12 attributes. Out of these attributes, 10 are considered relevant for the prediction task. These attributes provide valuable patient information, including their identification number, age, gender, hypertension, marital status, occupation, residence type, presence of heart disease, average glucose level, BMI, smoking habits, and stroke status.

A detailed examination of stroke occurrences concerning different features is presented in Fig. 1, with sub-figures. In sub-figure (Fig. 1a), it is visible that there is a slight increase in the number of strokes among females when compared to males. Moving on to sub-figure (Fig. 1b), a rising trend in stroke cases is observed as individuals age, with the highest incidence observed around the age of 80. Sub-figure (Fig. 1c) reveals that individuals with heart disease are more vulnerable to experiencing strokes. Marital status is explored in sub-figure (Fig. 1d), which suggests that married individuals may have a slightly higher incidence of strokes than unmarried individuals. The comparison between stroke occurrences in urban and rural areas is depicted in sub-figure (Fig. 1e), indicating no significant difference between these groups regarding stroke risk. In sub-figure (Fig. 1f), the relationship between average glucose levels and stroke risk is illustrated. It shows that individuals with average glucose levels falling within 60–120 and 190–230 are at an increased risk of experiencing strokes. Hypertension is emphasized in sub-figure (Fig. 1g). It demonstrates a higher incidence of strokes among individuals diagnosed with hypertension.

The relationship between BMI and stroke occurrence is examined in sub-figure (Fig. 1h). It reveals that individuals with a BMI ranging from 20 to 40 are more prone to strokes. Smoking habits are examined in sub-figure

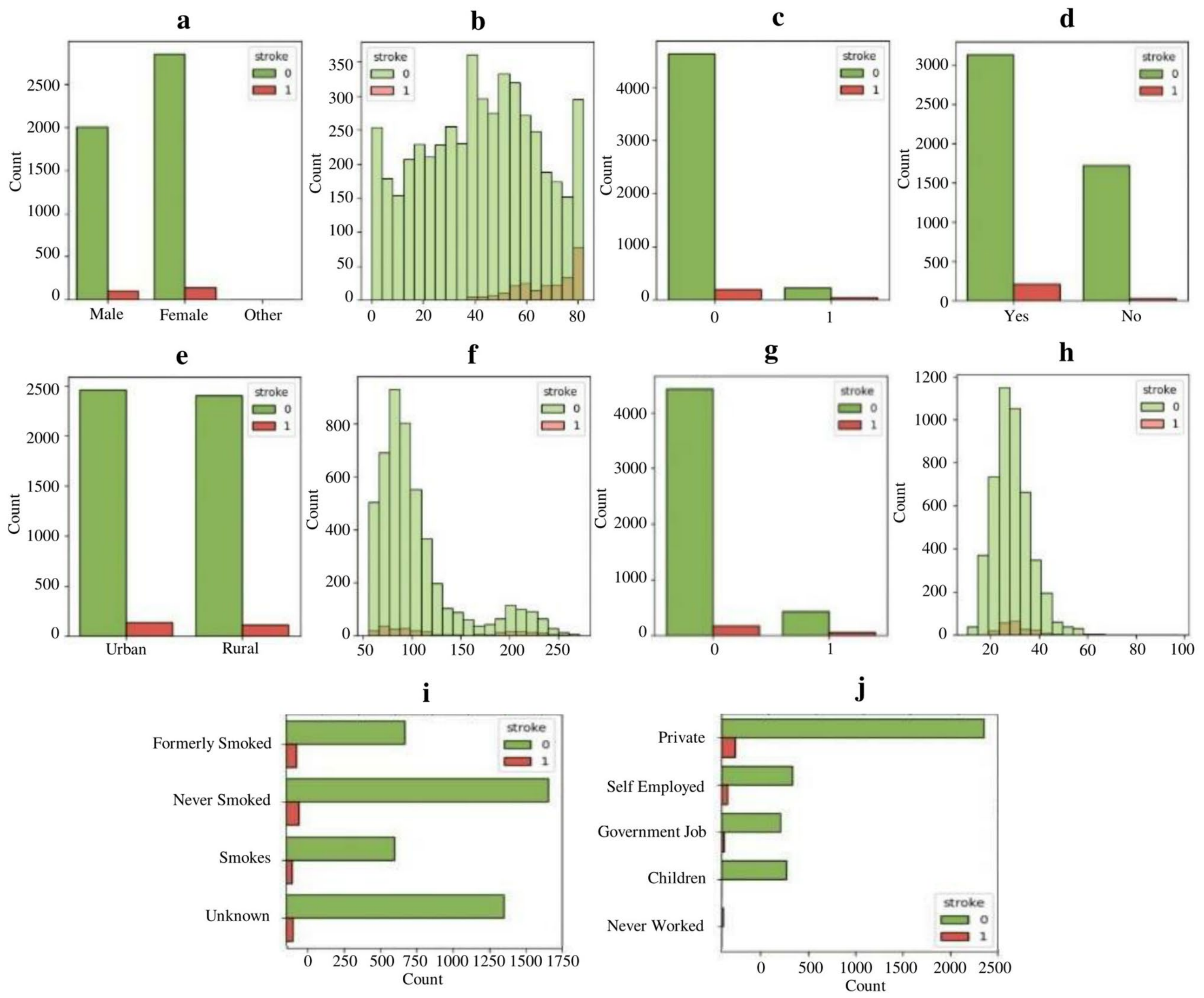

Figure 1. Distribution of features concerning stroke occurrence. (**a**) through (**j**) present diverse aspects of stroke occurrences, revealing nuanced patterns. (**a**) and (**b**) demonstrate gender and age-related trends. (**c**) associates strokes with heart disease, while (**d**) suggests marital status correlations. (**e**) explores urban–rural disparities. (**f**) and (**g**) show links to average glucose levels and hypertension. (**h**) relates BMI levels to stroke incidence. (**i**) emphasizes the role of smoking history, and (**j**) explores potential occupational influences on stroke likelihood.

(Fig. 1i), where it is observed that former or never smokers are more likely to suffer from strokes than current smokers. This finding highlights the importance of considering smoking history when assessing an individual's stroke risk. Lastly, shifting the focus to occupation, sub-figure (Fig. 1j) indicates that individuals working in private or self-employed sectors may have a greater likelihood of experiencing strokes compared to those in other occupations. This observation may be attributed to various factors such as stress levels, working conditions, and lifestyle differences among different occupational groups. Overall, the comprehensive analysis of stroke occurrences concerning different features provides valuable insights into the dataset and aids in understanding the factors contributing to stroke risk.

The dataset used in this research contains three numerical features: average glucose level, BMI, and age, while the remaining features are categorical. To assess the presence of outliers in the numerical features, box plots have been constructed and displayed in Fig. 2. The plots illustrate a notable presence of outliers in the average glucose level (Fig. 2b) and BMI metrics (Fig. 2c), emphasizing the need for meticulous data preprocessing. As depicted in Fig. 3, the distributions of individual numerical attributes diverge notably between those with and without a stroke. The discernible non-uniformity in these distributions underscores the importance of these features as promising indicators for stroke prediction.

A more in-depth exploration of the mentioned numerical attributes holds the promise of unravelling their influence on stroke prediction, thus offering invaluable insights to enhance the accuracy and efficacy of the predictive models. To preprocess the data, outliers are removed using the robust scaler method and apply standard scaling for consistent feature ranges. One-hot encoding is also utilized to convert categorical variables into binary values. Figure 4 illustrates the visual representation of encoded features correlation, offering valuable insights

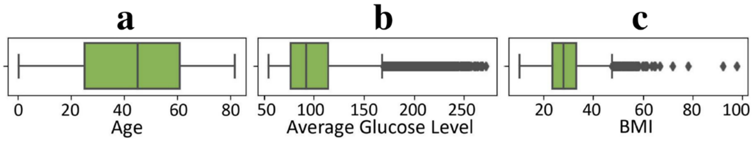

Figure 2. Box plots of numerical features to detect outliers. (**a**), (**b**) and (**c**) presents the boxt plots for age, BMI, and average glucose level to assess the presence of outliers.

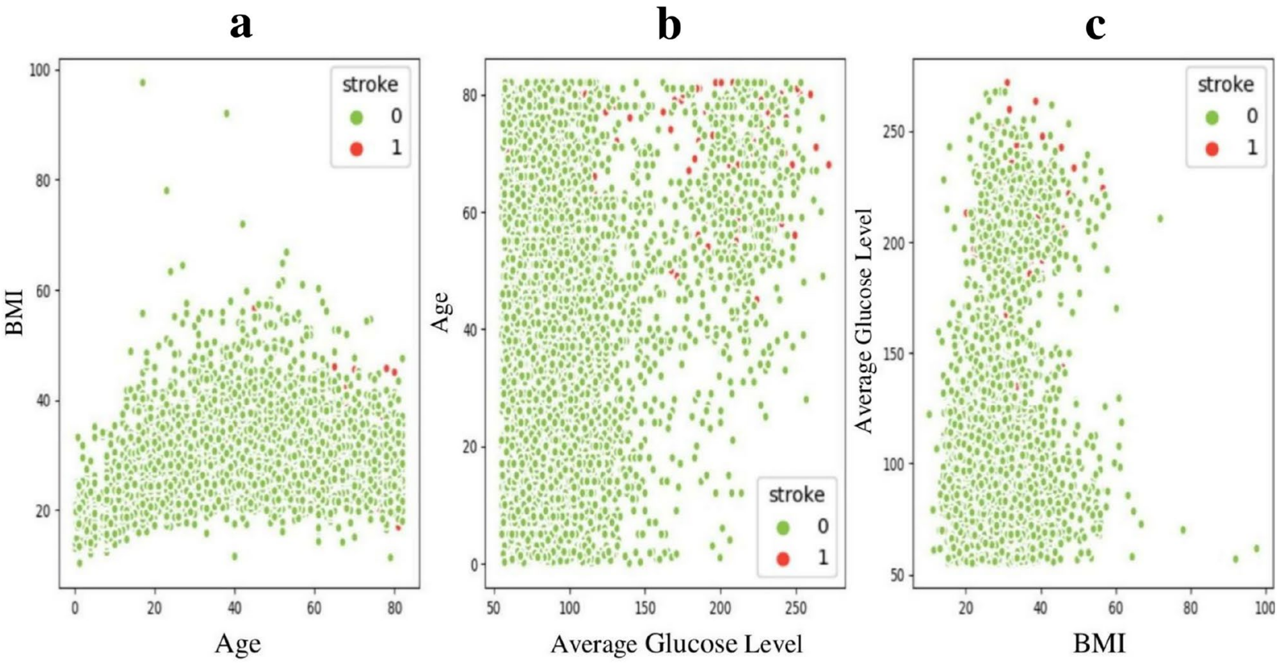

Figure 3. Distribution of numerical attributes with stroke and each other. (**a**), (**b**) and (**c**) presents the distribution plots for age, BMI, and average glucose level against eachother based on stroke occurences.

into the relationships between variables. This analysis helps to uncover significant associations and dependencies among the features, enhancing the understanding of the underlying patterns and dynamics within the dataset.

Missing data

The dataset will now be analyzed to predict stroke while inspecting for any missing values. The sparsity/nullity matrix for the dataset is shown in Fig. 5. It can be observed that the BMI feature has some missing values that need to be handled before proceeding with the analysis. As illustrated in Table 1, the dataset overview provides insights into the total number of stroke cases and the count of entries with missing BMI values. The table showcases that out of the 5110 total cases, there are 201 cases with missing BMI values. Consequently, if rows with missing values were dropped, there would be a data loss percentage of 3.93%, making it almost 4% of the dataset which can lead to loss of valuable insights present in it.

The naive response to missing values would be removing all those rows. But to avoid data loss through listwise deletion, this study will use imputation techniques to fill in the missing values. Imputation is a method of replacing missing data with an appropriate approximation based on available information. However, if not chosen carefully, imputation can introduce assumptions or biases. Therefore, this study explores well-established techniques specifically chosen for their ability to mitigate these potential issues. Later, the performance of these imputation techniques will be compared with the standard approach of dropping incomplete observations. This study explores three different imputation techniques to maximize data utility while maintaining data integrity and minimizing bias.

Mean ımputation

A commonly employed technique in data preprocessing involves replacing missing values in a dataset with the mean value of the respective variables within the feature column. This approach helps maintain data integrity and ensures the resulting dataset is complete and ready for analysis. Imputing missing values with the mean minimizes the impact of incomplete data on subsequent analyses. Additionally, this method enables us to use as much available information as possible, contributing to more accurate and robust results.

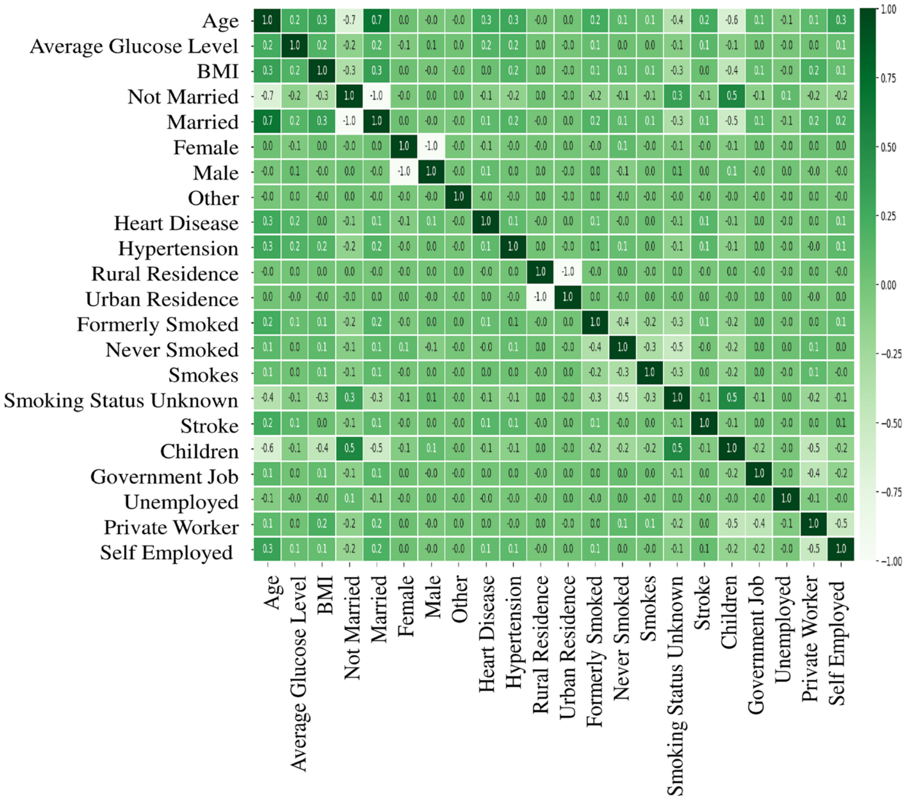

Figure 4. Features correlation heatmap for the dataset. Color intensity indicates the strength and direction of correlations, aiding in the identification of potential patterns and dependencies in the data.

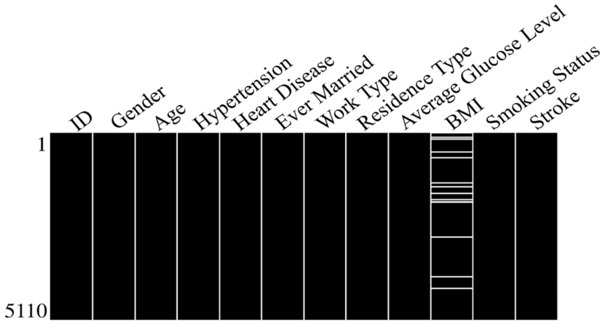

Figure 5. Sparsity matrix for the dataset. The empty spaces found in the corresponding column signify the presence of missing data values for the specific feature.

Total cases	Missing BMI values in cases	Data loss percentage due to dropping rows with missing values
5110	201	3.93%

Table 1. Evaluating the absence of data in datasets.

Multivariate ımputation using chained equations (MICE)

It is an advanced approach that surpasses single imputations. It employs multiple imputations, allowing a more robust estimation of missing values. The process involves a sequential regression technique, where each variable's missing values are estimated using information from other variables that have complete data. MICE significantly improves the accuracy and reliability of imputations, providing a comprehensive solution for handling missing values in datasets.

Age group-based BMI ımputation

To enhance the analysis, the individuals are classified into four age groups: 0–20, 21–40, 41–60, and 61–80. The strategy of imputing the mean BMI for each respective age group is employed to address missing values. This approach allows us to account for missing data while maintaining the integrity of the analysis. Additionally, this division into age groups enables a more nuanced understanding of the relationship between age and BMI, contributing to the overall accuracy and reliability of the findings.

Data ımbalance

After addressing the missing data, the focus is shifted to the data imbalance problem. There are more non-stroke case instances than stroke case instances, making the stroke case instances a minority class. This data imbalance poses a challenge in developing accurate predictive models and warrants the need for specialized techniques to handle class imbalance effectively. An overview of the class label populations in the dataset is presented in Table 2, revealing the presence of a minority class. This data imbalance can adversely affect model performance as the minority class is underrepresented. To mitigate the data imbalance issue, oversampling and undersampling techniques are commonly employed. Undersampling, which involves reducing the number of instances from the majority class, is not considered feasible. This approach may hinder the model's capacity to effectively learn patterns associated with stroke cases, potentially compromising its predictive accuracy. Oversampling is deemed feasible as it elevates the minority class distribution, so the predictive model performs well. However, if it is not implemented carefully, it may introduce the risk of overfitting. This study addresses this potential pitfall by investigating the efficacy of the models in predicting strokes using both balanced and imbalanced datasets. This approach will facilitates a rigorous evaluation of the predictive model's performance under diverse data conditions. To address this class imbalance issue, SMOTE is assessed as a potential remedy.

Synthetic minority oversampling technique

SMOTE is a widely recognized method for oversampling. It is employed to increase the representation of minority samples in a dataset. To understand how it operates. Consider a scenario where the training dataset consists of 's' samples and 'f' features. To enhance the representation of the minority class in the dataset, the methodology employed is known as oversampling. This technique involves selecting a sample from the minority class and identifying its k nearest neighbours in the feature space. Subsequently, it generates a new synthetic data point by combining the original data point with one of its closest neighbours. This combination is achieved by scaling the vector connecting the two points by a random number 'x' ranging from 0 to 1. Incorporating this synthetic data point into the existing dataset effectively addresses the class imbalance issue and generates a fresh, augmented data point. This process helps balance the classes and improve the overall representation of the minority class[42].

Data modelling

A robust data modelling approach is essential to effectively analyze and predict stroke occurrences, encompassing raw data's systematic transformation and organization into a structured framework. Figure 6 depicts the data modelling pipeline utilized in this research, showcasing the various stages and methodologies employed. This pipeline enhances the data analysis and prediction approach's accuracy and efficiency.

This research employs various techniques for stroke prediction to handle missing and imbalanced data issues. The researchers utilize mean, MICE, and age group-wise BMI mean imputation methods to handle missing values. To tackle the data imbalance issue, SMOTE is used to increase the representation of the minority class labels by generating synthetic samples. Additionally, the outliers are removed using the robust scaler method and apply standard scaling to ensure consistent feature ranges. Categorical variables are transformed into binary values through one-hot encoding. The predictive models encompass a baseline model, and then followed by advanced models including TabNet, Logistic Regression with AGD (LR-AGD), Neural Network, Random Forest, Gradient Boosting, CatBoost, LightGBM, XGBoost, Balanced Bagging, and NGBoost. The dataset is divided into training and testing data using a 70:30 ratio. The models are evaluated using k-fold cross-validation on both balanced and imbalanced imputed datasets and on the augmented dataset, generating multiple analyses for each model. Also, the trained models are tested on testing data to assess their generalization performance. The performance of each model is also compared to the standard approach of dropping incomplete observations on the original dataset. Consequently, a Dense Stacking Ensemble (DSE) model is built upon previous models after fine-tuning, with the best-performing model as a meta-classifier. Finally, all models are ranked and analyzed, including the DSE model, using various performance metrics.

Augmentation of dataset

The primary objective is to identify the key factors contributing to stroke prediction. To accomplish this, the authors highlight essential features in Fig. 7 that demonstrate a positive impact on stroke prediction and their corresponding importance factors. This analysis provides valuable insights into the factors that play a vital role

Total cases	Stroke cases	Non-stroke cases	Percentage of minority stroke case samples
5110	249	4861	4.87%

Table 2. Evaluating the imbalance of data in the dataset.

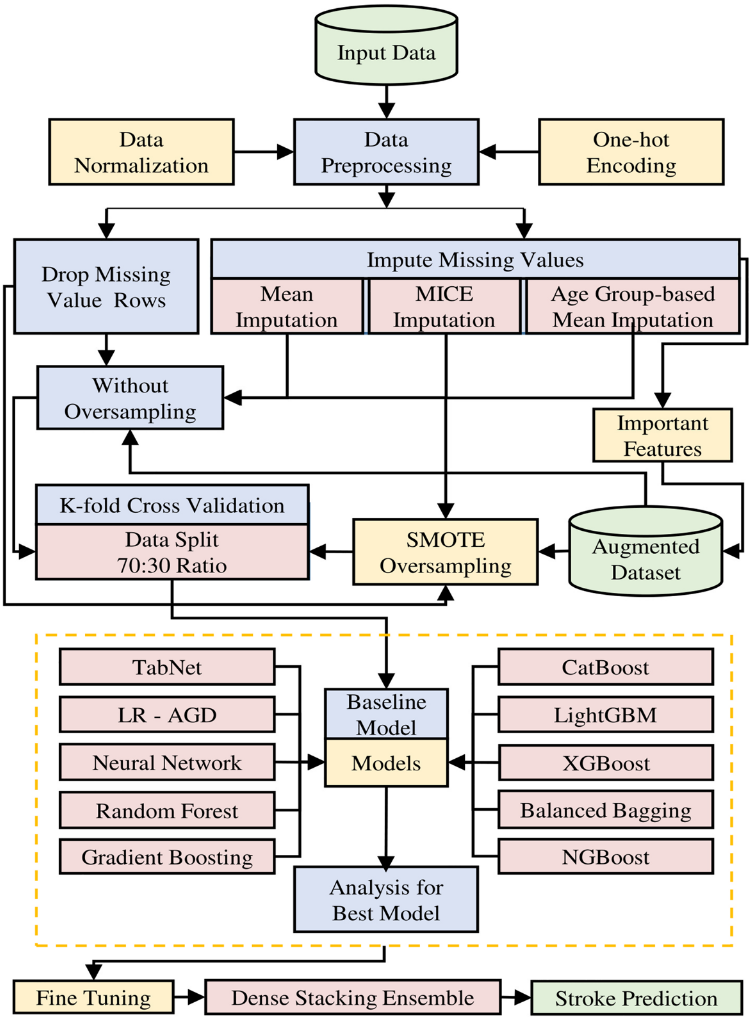

Figure 6. Pipeline for data modelling. Stroke prediction data modeling pipeline integrates techniques for missing and imbalanced data. Prediction models, from TabNet to NGBoost, undergo rigorous evaluation and testing, culminating in a Dense Stacking Ensemble (DSE) for enhanced and robust prediction results.

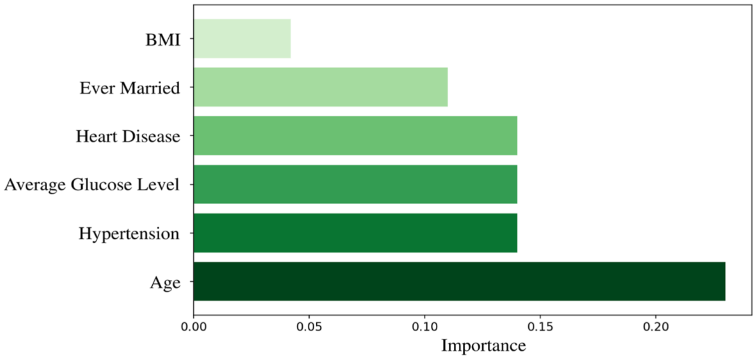

Figure 7. Features with positive importance factor for stroke prediction. Features of significant positive importance for stroke prediction include age, BMI, average glucose level, heart disease, hypertension, and marital status (ever-married) respectively.

in accurately predicting strokes. This study incorporates information from the previous three imputed datasets for dataset augmentation, resulting in a larger dataset with 10,421 distinct instances.

The augmented dataset includes age, BMI, average glucose level, heart disease, hypertension, ever-married, and stroke label features. Interestingly, the findings align with another previously conducted comprehensive study that used the same dataset[43], where the critical features identified for stroke prediction using the same dataset were the same, except for the inclusion of the "ever married" feature. It is noticed that the ever-married feature has a high frequency of stroke occurrences among those individuals who were or are married.

K-fold cross validation

To assess how well the models perform, it is required to divide the dataset into training, validation, and testing data. Since no separate "unlabeled" test dataset is available, this study adopts a tenfold validation method. This means that it divides the training dataset into K parts. During each iteration, a single part is designated as the validation dataset out of the K parts obtained by dividing the training dataset. The remaining parts are utilized for training the model. This process is repeated K times, allowing for a comprehensive evaluation of the model's performance. Performance metrics are recorded for each validation set. After all iterations, the metrics are averaged across all K iterations, ensuring that each bin served as a validation dataset at least once[44].

Machine learning algorithms

The main objective is to develop an accurate and robust predictive model for stroke prediction. The authors begin by using a baseline model to establish a reference point for model performance. They then investigate different advanced classification models to determine the accuracy of these models in predicting stroke. After thorough fine-tuning, the authors construct a robust DSE model that leverages the best-performing model as a meta-classifier. The upcoming subsections will delve into the classification models, which are also used as base models, as well as the architecture of the DSE model.

Baseline model

This study employs Logistic Regression as the baseline model for stroke prediction. It is a statistical technique widely used for binary classification tasks. It estimates the probability of a binary outcome based on input features using the logistic function as given in Eq. (1).

$$g(\mathrm{z}) = \frac{1}{1 + e^{-z}}, \tag{1}$$

where z represents the linear combination of input features weighted by corresponding coefficients. This model serves as an initial benchmark for evaluating the performance of more advanced classification models.

Advanced classification models

In the quest for creating a strong stroke prediction model, a variety of advanced contemporary classification models are carefully examined without fine-tuning and put to use. These models serve a dual purpose: first, they undergo rigorous evaluation for predictive accuracy, and second, they constitute the core elements of the DSE model, which employs a layered and efficient approach to predicting strokes.

TabNet

TabNet is a supervised machine learning algorithm that operates on tabular data and employs a neural network architecture with attention-based feature selection and sequential decision steps. It is designed to handle structured data and can effectively capture complex relationships between input features to make accurate predictions for stroke classification.

$$\hat{y} = \mathrm{TabNet}(X; \theta), \tag{2}$$

$$\theta = arg\ minL\left(y, \hat{y}\right). \tag{3}$$

In Eq. (2) and (3), let X be the input feature matrix, y be the binary target variable (0 or 1) representing the stroke, and $\hat{y}$ be the predicted stroke probabilities. The TabNet algorithm aims to find the optimal parameters θ that minimize the loss function L.

Logistic regression with AGD

Logistic Regression models the relationship between input variables and binary output. It utilizes the logistic function to estimate the probability of the outcome, making it suitable for binary classification tasks like stroke prediction. In this study, the logistic regression model is trained efficiently using the accelerated gradient descent (AGD) optimization technique. The model is limited to 100 maximum iterations during training.

$$\hat{y} = \frac{1}{1 + e^{-(X \times \beta)}}. \tag{4}$$

In logistic regression with AGD, the model estimates the probability of stroke $\hat{y}$ given the input features X using the logistic function, where β represents the model's coefficients as given in Eq. (4).

Neural network

Neural network is a powerful machine learning model that consists of interconnected nodes or "neurons" organized in layers. It is capable of learning complex patterns from data and making non-linear predictions. The neural network used in this research has five hidden layers with 24, 36, 48, 36, and 24 neurons, respectively. It is trained to recognize significant features related to stroke prediction and make accurate decisions based on them.

$$\widehat{y} = \sigma(X \times W_{out} + b_{out}). \tag{5}$$

In Eq. (5), the neural network involves a series of calculations with weight matrices (W) biases (b), and activation functions (σ) in each layer. The output layer uses the sigmoid activation function to obtain the predicted stroke probabilities ($\widehat{y}$). The W_{out} and b_{out} are the weight matrix and bias vector of the output layer, respectively.

$$\theta = arg\ min\ L\left(y,\widehat{y}\right), \tag{6}$$

where in Eq. (6), the neural network aims to minimize the loss function L during training.

Random forest

The Random Forest model builds a collection of decision trees and combines their predictions to make final predictions. It utilizes random feature selection and bootstrapping to create diverse tree models. The number of estimators is denoted as $N_{\mathrm{RF}} = 100$.

Gradient boosting

The Gradient Boosting model is a powerful predictive model that utilizes a combination of weak prediction models, commonly decision trees, to generate accurate predictions. This model iteratively enhances its predictions by fitting new models to the residuals of previous models. In this study, $N_{\mathrm{GB}} = 100$ estimators are employed to optimize the performance of the Gradient Boosting model.

CatBoost

The category Boosting algorithm is specifically designed for categorical data. It utilizes gradient boosting and implements novel techniques to handle categorical variables effectively. The number of estimators for this model is $N_{\mathrm{CB}} = 100$.

LightGBM

A light Gradient Boosting Machine is a gradient boosting framework that aims to provide high efficiency and speed. It uses a histogram-based approach for gradient boosting and incorporates features like leaf-wise tree growth and data parallelism. $N_{\mathrm{LGBM}} = 100$ estimators are used.

XGBoost

Extreme Gradient Boosting is a gradient boosting algorithm known for its scalability and performance. It combines multiple weak prediction models and employs regularization techniques to prevent overfitting. For this model, $N_{\mathrm{XGB}} = 100$ estimators are used.

Balanced bagging

Balanced Bagging is an ensemble learning algorithm that combines multiple classifiers by training them on different subsets of the original dataset. It explicitly addresses class imbalance issues by using sampling techniques to balance the class distribution. Five Random Forest Classifiers as base estimators are used.

NGBoost

Natural gradient boosting is a gradient boosting algorithm that focuses on probabilistic prediction and uncertainty estimation. It utilizes natural gradient boosting and incorporates Bayesian methods for improved model calibration. For this model, $N_{\mathrm{NGB}} = 100$ estimators and a learning rate of 0.01 are used.

Dense stacking ensemble model

The cornerstone for robust stroke prediction system is the DSE model, as its high-level architecture is visually depicted in Fig. 8. The DSE model is meticulously crafted to optimize predictive accuracy and robustness. The DSE architecture integrates a range of fine-tuned classification models, each playing a vital role as a base model.

Within the this model, three distinct approaches are employed, with each utilizing the best-performing model as a meta-classifier. These approaches are strategically integrated as base models in the final DSE model to enhance predictive accuracy and reliability. These approaches are explained in the following.

Voting ensemble

The Voting ensemble approach operates by collecting predictions from multiple base models and making a collective prediction based on the most popular choice. This approach leverages the best-performing model as the meta-classifier, which means it gives more weight to the predictions of this model. By combining the insights from various models, the Voting ensemble aims to maximize the overall predictive power of the DSE model. It's like having a panel of experts vote on the most likely outcome, with the best expert's opinion carrying the most weight[45].

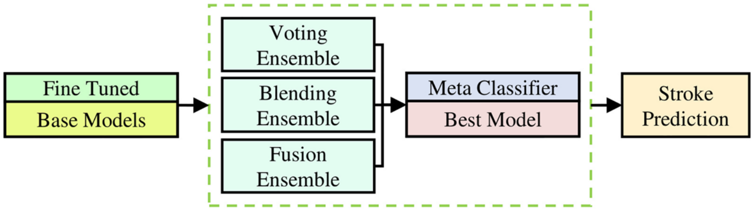

Figure 8. Architecture of the dense stacking ensemble model. The dense stacking ensemble integrates fine-tuned base models using three distinct approaches, each incorporating the best-performing model as a meta-classifier.

Blending ensemble

In the Blending ensemble approach, a meta-classifier is trained using the predictions made by the base models. The best-performing model takes on the role of the meta-classifier in this approach. This model is skilled at blending the predictions of other models in a way that optimizes their collective predictive strength. The Blending ensemble essentially learns how to combine the different model outputs best, capitalizing on the unique strengths of each model to enhance the overall predictive accuracy of the DSE model[46].

Fusion ensemble

The harmonious integration of both the base models and the best-performing model characterizes the Fusion ensemble approach. It doesn't just use the best-performing model as the meta-classifier; it collaboratively combines the strengths and insights of all models in a synergistic manner. This approach creates a final predictive model that benefits from the diverse perspectives and capabilities of the base models, thus producing a more robust and accurate prediction within the DSE framework[47]. It's like bringing together a team of experts to solve a complex problem, with each expert contributing their unique insights and skills.

Results

In this section, a comprehensive analysis of results, along with various score plots, is provided. The datasets are generated using Mean, MICE, and Age Group-based imputation techniques to address missing values. The analysis also encompasses the score plots for the original dataset, in which missing values are managed through list-wise deletion. Furthermore, the results for the augmented dataset are presented in a similar manner. The results are presented in three subsections: The first and second subsections focuse on the performance of various classification models, with a particular emphasis on comparing their effectiveness against a baseline model. The second subsection is about the DSE model results.

To gauge the effectiveness of the models, a diverse set of metrics is employed. These metrics encompass the confusion matrix, which includes true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values for actual and predicted data. The definition of these confusion matrix parameters can be found in Table 3. Additional metrics, such as accuracy, precision, recall, F1 score, and AUC are utilized to provide a comprehensive evaluation of classifier performance[48,49]. Table 4 explains these metrics briefly. The findings highlight the effectiveness of the different imputation techniques in handling missing data and showcase the impact on

Confusion matrix parameter	Description
True positive (TP)	The number of stroke case instances correctly predicted as stroke cases
True negative (TN)	The number of non-stroke case instances correctly predicted as non-stroke cases
False positive (FP)	The number of non-stroke case instances incorrectly predicted as stroke cases
False negative (FN)	The number of stroke case instances incorrectly predicted as non-stroke cases

Table 3. Definition of confusion matrix parameters.

Performance metric	Description	Formula
Accuracy	The number of correct predictions out of the total	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	The number of TP predictions out of the total positives	$\frac{TP}{TP+FP}$
Recall	The number of TP predictions out of the total actual positives	$\frac{TP}{TP+FN}$
F1 score	The mean of precision and recall as a balanced measure	$\frac{2\times Precision\times Recall}{Precision+Recall}$
AUC	The capacity of the model to differentiate between correct and incorrect predictions	*TP* and *FP* rate

Table 4. Definition of performance metrics.

model accuracy. The analysis of the augmented dataset demonstrates the improvement achieved by incorporating essential features. Overall, this section contributes to a comprehensive understanding of the various model's performance and insights for future research and model development.

Results of baseline model

The performance of the baseline model varied notably between the imbalanced and balanced datasets, as shown in Fig. 9. On the imbalanced dataset, the highest F1 score achieved was only 23.09%, which is considerably lower than the F1 score of 87.04% observed on the balanced datasets. This discrepancy is consistent across all other metrics, including accuracy, precision, recall, and AUC. In contrast, the baseline model exhibited notably higher performance on the balanced dataset, both in terms of the original data and the imputed datasets. For instance, on the original dataset, the baseline model attained the highest accuracy of 87.12%, precision of 87.65%, recall of 86.44%, F1 score of 87.04%, and AUC of 87.12%. This level of performance was closely mirrored in the imputed datasets, with the MICE-imputed dataset showing an accuracy of 85.54%, precision of 85.95%, recall of 85.44%, F1 score of 85.46%, and AUC of 85.54%. These findings suggest that either dropping rows with missing values or employing imputation techniques can significantly enhance the usability of the dataset, likely due to the increased availability of data for training the models.

Results of advanced classification models

In this section, we present the outcomes of our rigorous evaluation of advanced classification models. These models have been extensively assessed to provide insights into their predictive performance for stroke occurrences.

Imbalanced and balanced ımputed datasets results

MICE imputation yields slightly better results than Mean and Age Group-based imputation among the imbalanced imputed datasets. The top-performing model across all imputed datasets is LR-AGD, followed by NGBoost and Balanced Bagging. LR-AGD achieves a k-fold mean accuracy of 94.94%. The precision, recall, and F1 score are 95.20%, 94.94%, and 92.50%, respectively. Figure 10 shows the k-fold mean accuracy for all models on the imbalanced imputed datasets.

The Age Group-based balanced imputed dataset performs slightly better than the Mean and MICE imputed balanced datasets. XGBoost is the top-performing model, followed closely by LightGBM and Random Forest. The k-folds mean accuracy for the top-performing model is 93.48%, with precision, recall, and F1 score of 94.03%, 93.77%, and 93.76%, respectively. Figure 11 presents the k-fold mean accuracy of all models on balanced imputed datasets.

On the imbalanced MICE imputed testing dataset, LR-AGD achieves an accuracy of 96.28%, precision of 100%, recall of 14.93%, and F1 score of 25.97%. Moreover, for XGBoost on balanced Age Group-Based imputed testing dataset, the testing accuracy, precision, recall, and F1 score are 96.37%, 96.62%, 96.09%, and 96.35% respectively. The confusion matrices for the LR-AGD and XGBoost model on the respective testing datasets are displayed in Fig. 12. The matrix provides a visual representation of the model's performance. It allows for a comprehensive analysis of the model's accuracy and the distribution of correct and incorrect predictions across different classes.

The LR-AGD model consistently displayed the highest precision for the imbalanced datasets, with a value of 95.19%. It showcased its ability to predict positive instances accurately. On the other hand, when considering the balanced datasets, the XGBoost, LightGBM, and Random Forest models emerged as the top performers in terms of precision. The XGBoost model demonstrated the highest precision, ranging from 95.56% to 95.68%, followed closely by the LightGBM and Random Forest models with 95.53% and 94.96%, respectively. The k-fold mean precision of models on all imputed datasets is displayed in Fig. 13.

NGBoost, Balanced Bagging, and LR-AGD models consistently show high recall values, ranging from 94.91% to 94.94%, across different imputation techniques for the imbalanced datasets. In the case of balanced datasets,

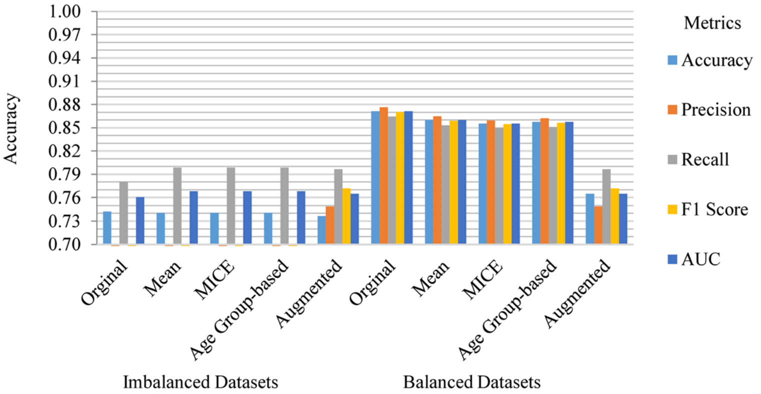

Figure 9. Baseline model performance across various datasets. The performance on imbalanced datasets is substantially lower than on the balanced datasets.

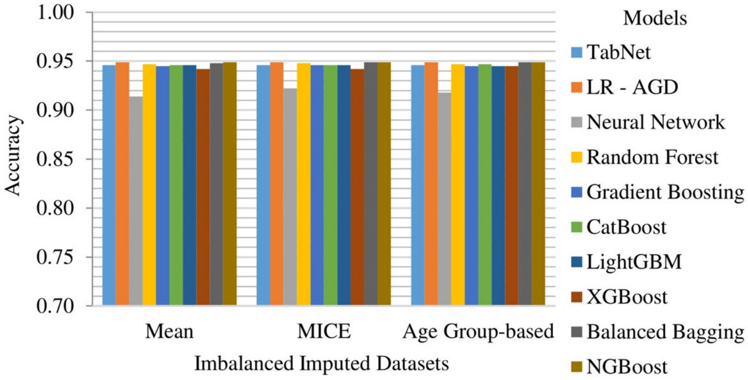

Figure 10. Models k-fold mean accuracy on imbalanced imputed datasets. The imputed imbalanced datasets are created using three different techniques namely as man, MICE, and age group-based.

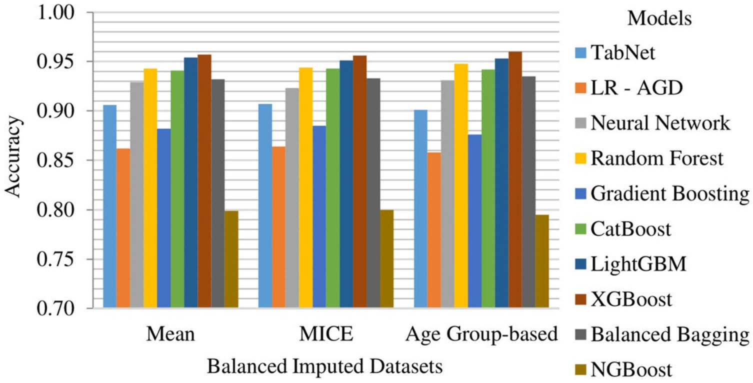

Figure 11. Models k-fold mean accuracy on balanced imputed datasets. The imputed balanced datasets are created using three different techniques namely as man, MICE, and age group-based.

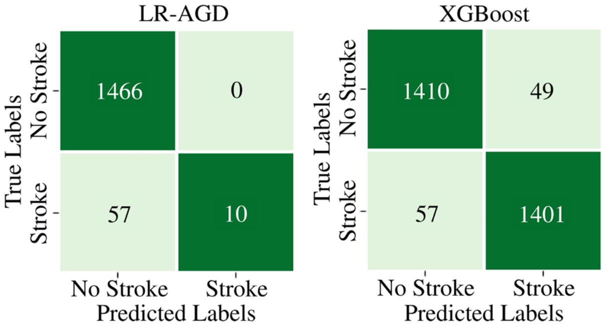

Figure 12. Confusion matrices of LR-AGD and XGBoost on imputed datasets. Confusion matrix illustrating the performance of LR-AGD (right) and XGBoost (left) models in stroke case classification on imbalanced and balanced imputed datasets, respectively.

XGBoost, LightGBM, and Random Forest models exhibit higher recall values of 95.90%, 95.53%, and 94.84%, respectively. These results indicate that the models are generally effective in capturing actual positive instances and correctly identifying them. The k-fold mean recall of models on all imputed datasets is displayed in Fig. 14.

Among the models evaluated, surprisingly, the Random Forest model achieved the highest F1 score of 94.72% on the imbalanced MICE imputed dataset and consistently performed well across all imputation techniques. XGBoost and LightGBM also performed well. When considering the balanced datasets, XGBoost, LightGBM, and Random Forest models exhibited the highest F1 scores, with values of 95.90%, 95.53% and 94.82%, respectively. Overall, these results indicate that XGBoost, LightGBM, and Random Forest models are promising models for

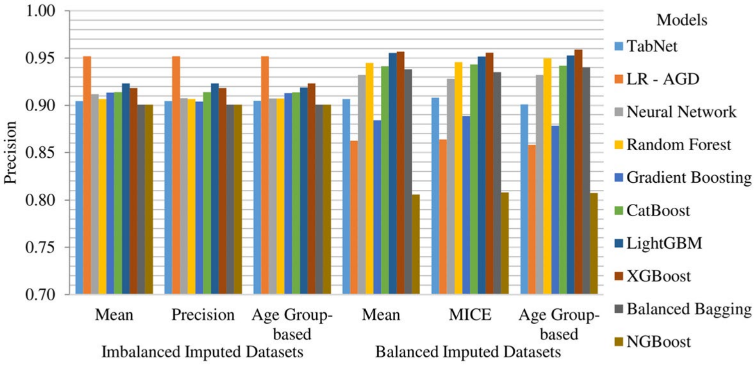

Figure 13. Models k-fold mean precision on all imputed datasets. The imputed datasets are created using three different techniques namely as man, MICE, and age group-based and divided into two categories named as imbalanced and balanced.

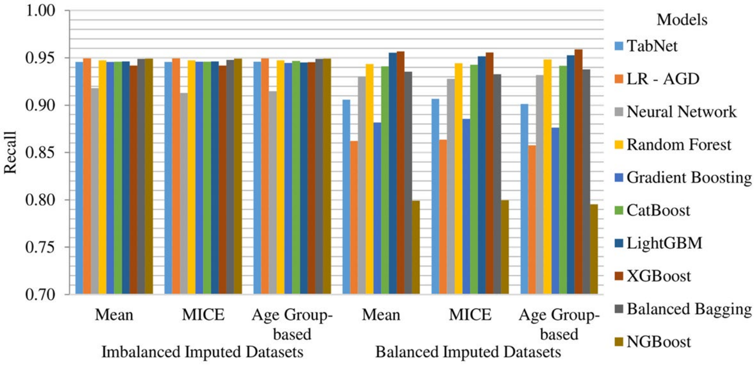

Figure 14. Models k-fold mean recall on all imputed datasets. The imputed datasets are created using three different techniques namely as man, MICE, and age group-based and divided into two categories named as imbalanced and balanced.

stroke prediction, showcasing their ability to achieve accurate classifications across different dataset characteristics. The k-fold mean F1 score of models on all imputed datasets is displayed in Fig. 15.

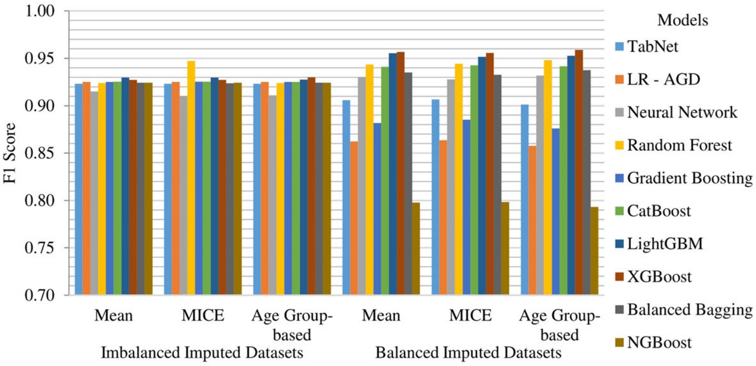

Figure 15. Models k-fold mean f1 score on all imputed datasets. The imputed datasets are created using three different techniques namely as man, MICE, and age group-based and divided into two categories named as imbalanced and balanced.

Imbalanced and balanced original datasets results

Among the models evaluated on the original dataset, which is created by removing the missing value rows, LR-AGD emerges as the top performer, closely followed by NGBoost and Balanced Bagging. LR-AGD achieves a k-fold mean accuracy of 95.46%, while NGBoost and Balanced Bagging achieve 95.46% and 95.43%, respectively. LR-AGD exhibits a precision, recall, and F1 score of 94.03%, 93.77%, and 93.76%, respectively. These exceptional results underscore the robustness and efficacy of LR-AGD in handling complex data scenarios. The k-fold mean accuracy for all models on imbalanced and balanced datasets is displayed in Fig. 16. For a balanced dataset with removed missing value rows, XGBoost emerges as the top-performing model, achieving a k-fold mean accuracy of 96.14%. It also exhibits high precision, recall, and F1 score, all at 96.14%. LightGBM and Random Forest follow closely behind in terms of second and third best models.

LR-AGD on imbalanced testing original dataset achieves the testing accuracy of 96.81%, precision of 87.50%, recall of 13.21%, and F1 score of 22.95%, respectively. When tested on the balanced original dataset, XGBoost maintains its superior performance with a testing accuracy of 98.33% and F1 score of 98.86%, while precision of 98.95%, and recall of 98.77%. The confusion matrices for the LR-AGD and XGBoost models on the respective testing datasets are shown in Fig. 17.

On the imbalanced original dataset, LR-AGD consistently achieved the highest precision and recall scores, with values of 93.42% and 95.46%, respectively. Regarding precision, the second and third best models are Neural Network and Balanced Bagging. NGBoost and Balanced Bagging models also showed impressive recall, the same as the LR-AGD. When considering the F1 score, which provides a balanced measure of precision and recall, all models achieved similar scores of around 93.3% except Neural Network. On the other hand, on the balanced datasets, XGBoost, LightGBM, and Random Forest consistently outperformed the other models regarding precision, recall, and F1 score, with values of 96.14%, 95.78%, and 95.44%, respectively.

Above mentioned results highlight the effectiveness of LR-AGD, XGBoost, LightGBM, and Random Forest in accurately classifying strokes across imbalanced and balanced datasets depending on the nature of the balance in the dataset, emphasizing their potential for stroke prediction applications. Furthermore, the k-fold mean precision, recall, and F1 score of these models on both imbalanced and balanced original datasets are visually depicted in Fig. 18.

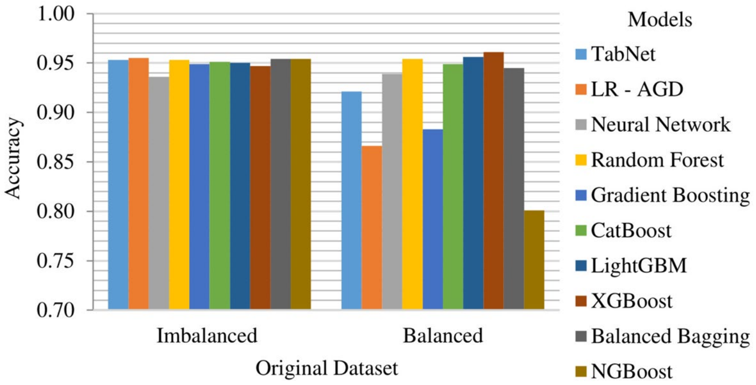

Figure 16. Models k-fold mean accuracy on original datasets. The dataset is categorized into two groups: imbalanced, reflecting its initial state, and balanced, achieved after employing oversampling technique.

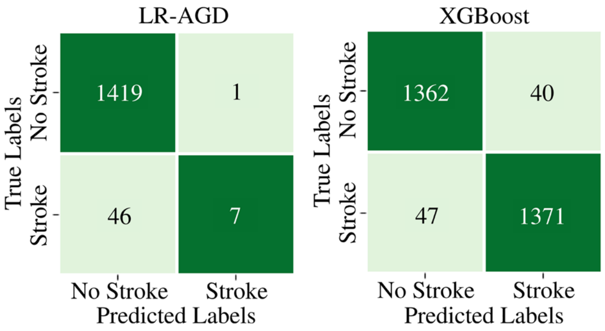

Figure 17. Confusion matrices of LR-AGD and XGBoost on original datasets. Confusion matrix illustrating the performance of LR-AGD (right) and XGBoost (left) models in stroke case classification on imbalanced and balanced original datasets, respectively.

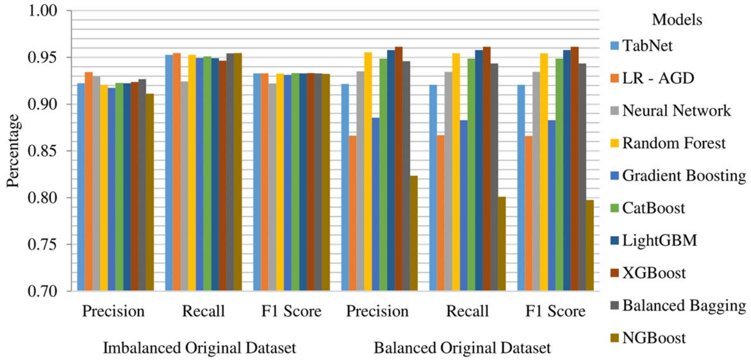

Figure 18. Models k-fold mean performance metrics on original datasets. The dataset is categorized into two groups: imbalanced, reflecting its initial state, and balanced, achieved after employing oversampling technique.

Imbalanced and balanced augmented datasets results

The Random Forest model demonstrates exceptional performance on the augmented dataset with imbalance, achieving a k-fold mean accuracy of 97.41%. It also exhibits high precision, recall, and F1 score, with values of 97.23%, 97.41%, and 97.14%, respectively. The remaining models display mean accuracies ranging from 95 to 96%. Figure 19 depicts the k-fold mean accuracy of all models on the augmented dataset, both imbalanced and balanced. The top performing model on the balanced augmented dataset is Random Forest, achieving a k-fold mean accuracy of 99.45% with a precision of 99.46% and 99.45% of recall and F1 score. It is closely followed by Balanced Bagging and XGBoost, with mean accuracies of 99.07% and 97.94%, respectively.

When evaluated on the imbalanced testing dataset, the Random Forest model maintains its strong performance with a testing accuracy of 97.57%, precision of 85.94%, recall of 65.48%, and F1 score of 74.32%. Random Forest also demonstrates impressive performance in testing on balanced dataset, with an accuracy of 99.61%, precision of 99.23%, recall of 100%, and an F1 score of 99.61%. The confusion matrices of the Random Forest model on imbalanced and balanced augmented testing datasets are depicted in Fig. 20.

The Random Forest model consistently demonstrates strong performance across all metrics on imbalanced and balanced augmented datasets. It achieves high precision scores of 97.23% on imbalanced data and 99.46% on balanced data, indicating its ability to identify true positive instances accurately. The Random Forest model also exhibits impressive recall scores of 97.41% on imbalanced data and 99.45% on balanced data, highlighting its capability to capture a high proportion of actual positive instances. In terms of F1 score, Random Forest Model achieves a balanced performance with scores of 97.14% on imbalanced data and 99.45% on balanced data. This indicates a harmonious balance between precision and recall, emphasizing its effectiveness in stroke prediction. The Balanced Bagging and XGBoost models also deliver competitive results across all metrics, showcasing their potential for accurate classification on balanced and imbalanced datasets. The k-fold mean precision, recall, and F1 score of models on all augmented datasets are shown in Fig. 21.

Results of dense stacking ensemble model

In the analysis of base model results, it becomes apparent that the MICE-imputed datasets produce marginally superior outcomes. Notably, the Random Forest model stands out as a top performer. Within the DSE model, the Random Forest model assumes the role of the meta-classifier, while the remaining models serve as base

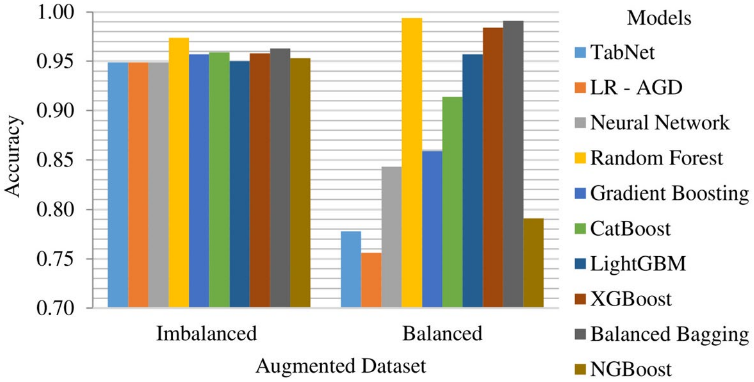

Figure 19. Models k-fold mean accuracy on augmented datasets. The dataset is categorized into two groups: imbalanced, and balanced, achieved after employing oversampling technique.

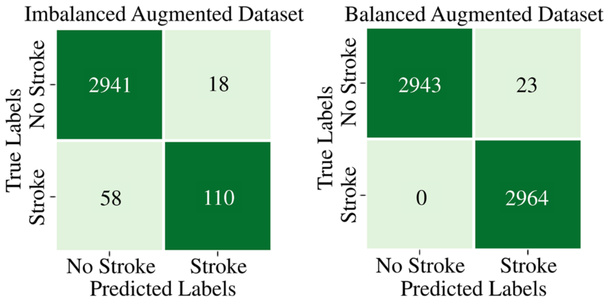

Figure 20. Confusion matrices of random forest on augmented datasets. Confusion matrix illustrating the performance of Random Forest model in stroke case classification on imbalanced (left) and balanced (right) augmented datasets, respectively.

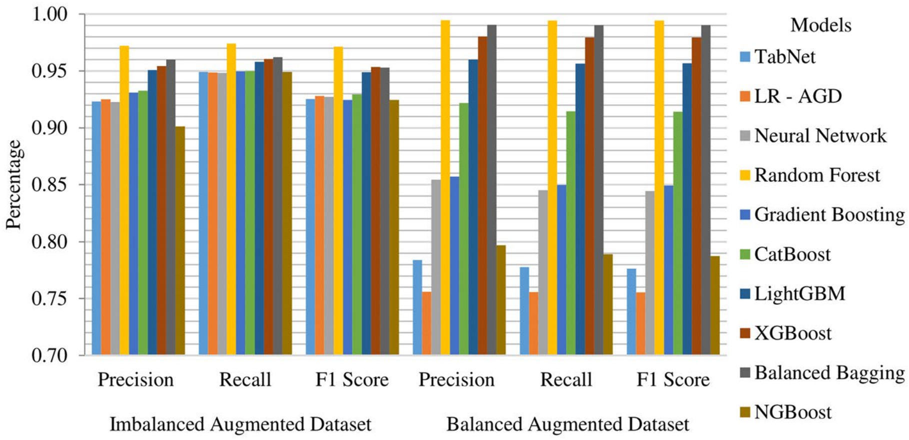

Figure 21. Models k-fold mean performance metrics on augmented datasets. The dataset is categorized into two groups: imbalanced, and balanced, achieved after employing oversampling technique.

models, highlighting the synergy derived from their collective strengths. The DSE model showcased remarkable performance on MICE-imputed datasets. For the imbalanced MICE-imputed datasets, the model yielded an accuracy of 96.13%, precision of 93.26%, recall of 96.18%, and an F1 score of 94.88%. Similarly, on balanced MICE-imputed datasets, the DSE model achieved an accuracy of 96.59%, with precision of 95.25%, recall of 96.27%, and F1 score of 95.79%. These results, also visualized in Fig. 22, highlight the robust performance of the DSE model when tested on imputed datasets.

The analysis of the AUC metric for the proposed DSE model reveals compelling insights into its predictive performance across different datasets. On the imbalanced dataset with MICE imputation, the DSE model achieves

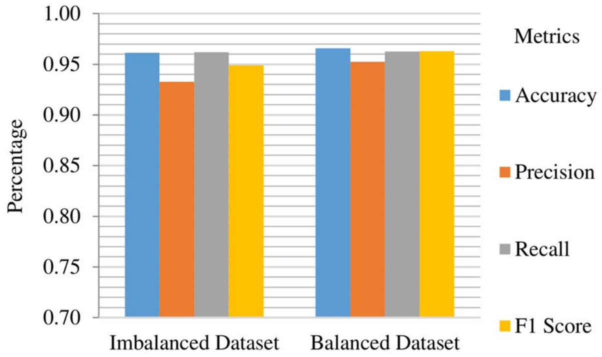

Figure 22. K-fold mean performance matrices of the proposed DSE model. The MICE imputed dataset is categorized into two groups: imbalanced, and balanced, achieved after employing oversampling technique.

an AUC of 83.94%, showcasing its ability to discern between positive and negative instances despite the data's skewed distribution. Conversely, on the balanced dataset, the DSE model excels even further, attaining an impressive AUC of 98.92%. This substantial increase in AUC on the balanced dataset underscores the model's enhanced discriminatory power and robustness when trained on a more representative and balanced data distribution. The significant performance improvement achieved by the DSE model on the balanced dataset compared to the imbalanced one is visually represented in Fig. 23.

Furthermore, on the imbalanced testing dataset, the model shows a testing accuracy of 99.15%, a precision of 84.93%, a recall of 98.88%, and an F1 score of 90.51%. For the balanced testing dataset, the model gives a testing accuracy of 97.19%, precision of 96.83%, recall of 97.38%, and an F1 score of 97.10%. The confusion matrices of the DSE model on imbalanced and balanced MICE-imputed testing datasets are depicted in Fig. 24.

The analysis of feature importance is conducted to ascertain the influential factors in stroke prediction using the proposed DSE model. The analysis of feature importance revealed distinct patterns between the imbalanced and balanced datasets, as visualized in Fig. 25. In both datasets, the top three features influencing stroke prediction were average glucose level, age, and BMI. However, notable differences were observed in their relative importance. In the imbalanced dataset, these top three features were relatively close in importance, with average glucose level slightly more influential than age and BMI. Conversely, in the balanced dataset, age emerged as the most important feature by a significant margin, followed by average glucose level and BMI. Additionally, the imbalanced dataset highlighted hypertension and heart disease as the 4th and 5th most important features, while the balanced dataset indicated that marital status (yes and no) played a more significant role in prediction. Interestingly, features such as work type (never worked and children) and gender (other) showed minimal contribution in both datasets, underscoring their limited impact on stroke prediction outcomes.

Discussion

While LR-AGD and XGBoost deliver accurate results with high accuracy, they both exhibit limitations. LR-AGD performs well when the data is imbalanced, but its performance significantly decreases when the dataset is balanced, behaving differently and yielding lower accuracy. Conversely, XGBoost performs exceptionally well on balanced datasets but struggles with imbalanced ones. Linear models excel in simpler, easily separable small data, while non-linear models perform better in complex and equally represented variables with intricate relationships. This highlights the importance of creating an appropriate model to handle these kinds of versatile dataset characteristics to yield optimal performance in stroke prediction. It is crucial to consider the balance between

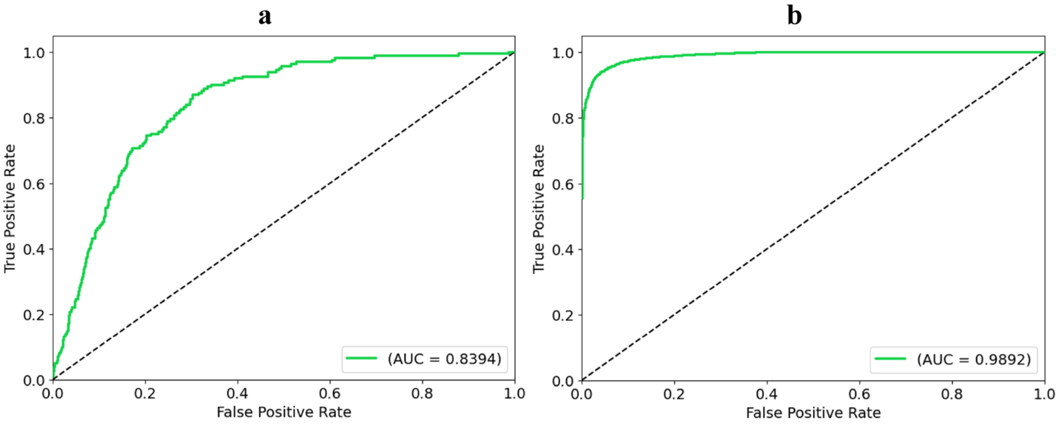

Figure 23. AUC results of the proposed DSE model. (**a**) represents the AUC results on imbalanced MICE imputed dataset and (**b**) on the balanced one.

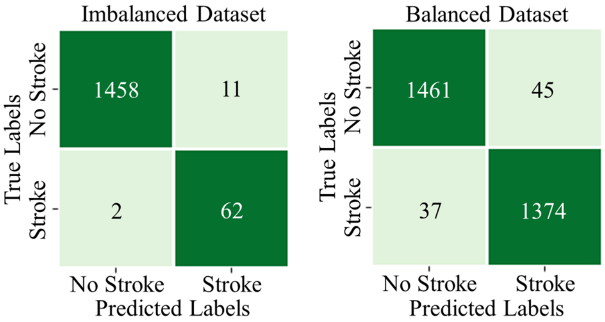

Figure 24. Confusion matrices of the proposed DSE model. Proposed DSE model's confusion matrices on MICE imputed balanced and imbalanced datasets.

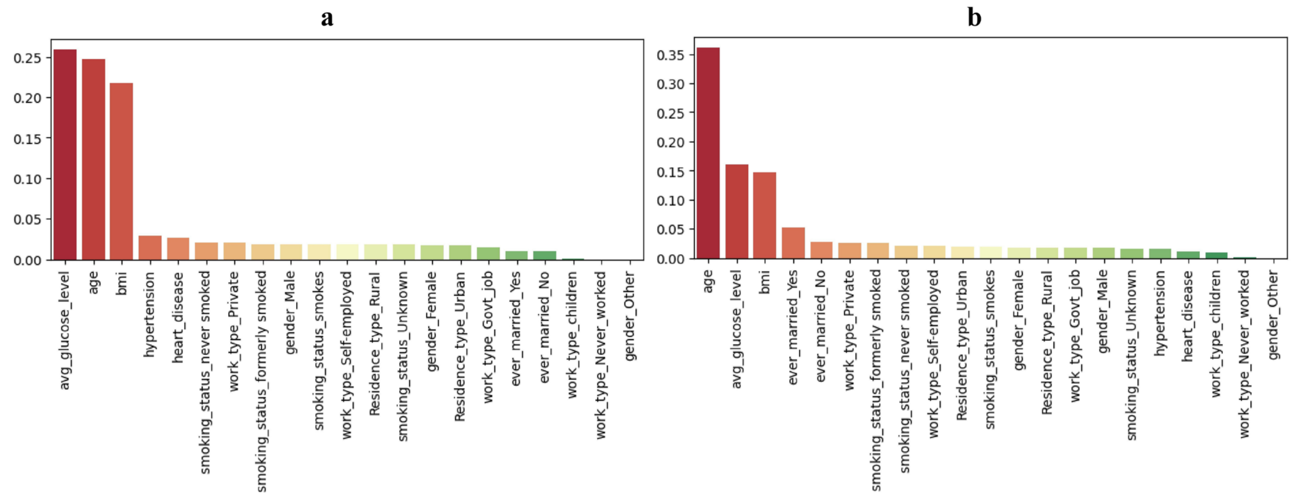

Figure 25. Feature importance comparison for the proposed DSE model. Feature importance graphs for imbalanced and balanced MICE-imputed datasets are displayed in (**a**) and (**b**) respectively.

precision and recall to make informed decisions regarding model selection. Additionally, further research and development are needed to address the limitations of LR-AGD and XGBoost to enhance their performance across various dataset scenarios.

However, when an augmented dataset is created, incorporating crucial factors significantly contributing to stroke prediction, Random Forest emerges as the superior model. It consistently outperforms other models on both imbalanced and balanced augmented datasets with a mean accuracy of 97.409% and 99.068%, respectively. Random Forest also gives consistent and around 95% accurate results for non-augmented datasets. Ultimately, the Random Forest is then used as a meta-classifier in the DSE model. Tables 5 and 6 provide a comprehensive summary of the mean accuracy of advanced classification models and the DSE model across all imbalanced and balanced datasets, highlighting the superior performance of the DSE model. The DSE model achieves far more superior results when other models are incorporated within it as base models and Random Forest as meta-classifiers. The DSE achieves the highest accuracy ranging above 96% across all types of datasets, making it the most feasible and robust model for stroke prediction on diverse datasets.

Additionally, Table 7 compares stroke prediction results from the previous recent studies that utilized the same dataset. This comparative analysis provides valuable insights into the top-performing DSE machine learning model's performance on imbalanced and balanced datasets, showcasing its respective accuracies. The table serves as a comprehensive reference for understanding the effectiveness of these models in stroke prediction. The study[27] shows that the minimal genetic folding (mGF) model achieves an accuracy of 83.2% on the balanced dataset. Another study[28] utilizes Logistic Regression and achieves an accuracy of 86.00%. Naive Bayes[29] achieves an accuracy of 82.00%. Random Forest[30] achieves an impressive accuracy of 94.46% on the imbalanced dataset, while Support Vector Machine[31] reaches an accuracy of 95.49%. Additionally, Random Forest is studied[32–34,36] with accuracies ranging from 95.50% to 96.00%. The proposed RXLM[35] model achieves an accuracy of 96.34% on the balanced dataset. K-nearest Neighbours[37] model achieves accuracies of 94.00% on the balanced dataset. In this study, the proposed DSE model achieves an impressive accuracy of 96.13% on the imbalanced imputed

	Imbalanced datasets				
		Imputed			
Model	Original	Mean	MICE	Age group-based	Augmented
TabNet	95.26%	94.58%	94.58%	94.60%	94.91%
LR-AGD	95.46%	94.94%	94.94%	94.94%	94.87%
Neural network	92.96%	91.40%	92.17%	91.81%	94.93%
Random forest	95.28%	94.72%	94.72%	94.74%	97.41%
Gradient boosting	94.96%	94.55%	94.60%	94.49%	94.97%
CatBoost	95.11%	94.60%	94.60%	94.66%	95.00%
LightGBM	94.91%	94.63%	94.63%	94.52%	95.80%
XGBoost	94.64%	94.19%	94.19%	94.55%	96.05%
Balanced bagging	95.43%	94.86%	94.94%	94.86%	96.28%
NGBoost	95.46%	94.91%	94.91%	94.91%	94.93%
Proposed DSE	96.52%	96.08%	96.13%	96.11%	98.16%

Table 5. Mean accuracies of models across all imbalanced datasets.

	Balanced datasets				
		Imputed			
Model	Original	Mean	MICE	Age group-based	Augmented
TabNet	92.08%	90.58%	90.70%	90.12%	77.77%
LR-AGD	86.60%	86.23%	86.36%	85.78%	75.57%
Neural network	93.89%	92.90%	92.70%	93.12%	84.29%
Random forest	94.44%	94.36%	94.43%	94.83%	99.45%
Gradient boosting	88.27%	88.18%	88.55%	87.53%	85.00%
CatBoost	94.85%	94.11%	94.28%	94.17%	91.46%
LightGBM	95.75%	95.53%	95.15%	95.27%	95.64%
XGBoost	96.14%	95.68%	95.96%	95.50%	97.94%
Balanced bagging	94.51%	93.18%	93.34%	93.48%	99.07%
NGBoost	80.09%	79.91%	79.97%	79.53%	78.91%
Proposed DSE	96.88%	96.57%	96.59%	96.56%	99.85%

Table 6. Mean accuracies of models across all balanced datasets.

	Imbalanced dataset	Balanced dataset
Model	Accuracy	Accuracy
Minimal genetic folding[27]	–	83.20%
Logistic regression[28]	–	86.00%
Naïve Bayes[29]	–	82.00%
Ensemble random forest[30]	94.46%	–
Support vector machine[31]	95.49%	–
Random forest[32]	95.50%	–
Random forest[33]	–	96.00%
Random forest[34]	–	96.00%
Proposed RXLM[35]	–	96.34%
Random forest[36]	–	96.00%
K-nearest neighbours[37]	–	94.00%
Proposed dense stacking ensemble (DSE)	96.13%	96.59%

Table 7. Comparison of stroke prediction results with previous studies on same dataset.

dataset and 96.59% on the balanced dataset. The notable distinction in the DSE model's performance can be attributed to its unique ability to harness the strengths of multiple base models through ensemble techniques. By employing a strategic combination of Voting, Blending, and Fusion ensembles, the DSE model maximizes predictive accuracy by leveraging the diverse perspectives and capabilities of each individual model. This sophisticated integration of ensemble methods enables the DSE model to outperform standalone models, as seen in previous studies. Overall, Table 7 provides a comprehensive overview of stroke prediction results, showcasing the performance of previously used models on imbalanced and balanced datasets along with the performance of the proposed DSE model.

In the domain of practical application, the DSE model exhibits a seamless integration into a real-life scenario, as demonstrated in Fig. 26. Users, whether they be individuals concerned about their health or medical professionals, can effortlessly input vital signs and demographic information through a user-friendly mobile or web application. This data is then securely transmitted to a cloud server where the pre-trained DSE model is deployed. The model processes the input information swiftly, with an average prediction time of 0.095 s per subject, showcasing its efficiency. Upon completion of the prediction, results are promptly relayed back to the user through the same cloud server, accessible via the mobile or web app. Crucially, if a subject is predicted to be at risk of a stroke, the system offers the option for immediate online consultation with a healthcare professional or assistance in locating a physical medical service through a third-party service. This innovative approach not only underscores the model's applicability in real-world scenarios but also highlights its potential to contribute significantly to proactive healthcare management.

Conclusion

Given the substantial global impact of strokes on mortality rates, there is an urgent need for robust and generalizable early prediction methods. While stroke prediction models are pivotal in pinpointing high-risk individuals, they face obstacles such as missing data and data imbalance. This study aims to create an improved predictive model for stroke prediction and evaluate its performance across various imbalanced and balanced datasets. The comprehensive analysis of various advanced machine learning models for stroke prediction that are presented

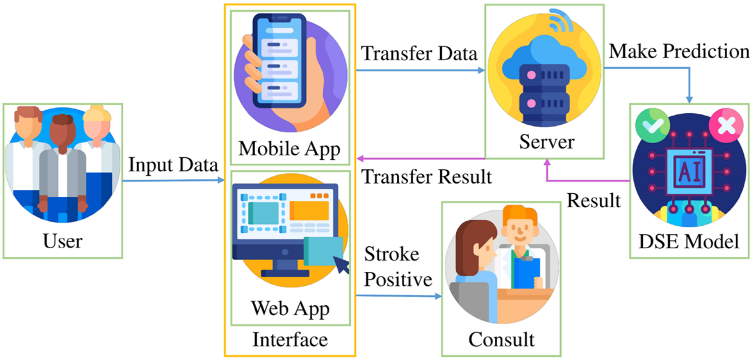

Figure 26. Integration of proposed DSE model in real-life scenario. The DSE model quickly processes vital signs on a user-friendly app, offering timely stroke risk predictions via a cloud server.

in this research paper sheds light on the efficacy of different techniques and models in handling missing data and data imbalance. The study reveals that most significant factors for stroke prediction are age, BMI, average glucose level, heart disease, hypertension, and ever-married status. Subsequently, an augmented dataset is created to incorporate these essential features, with the goal of enhancing the accuracy of stroke prediction models. The study uses an extensive range of advanced models such as TabNet, Logistic Regression, Neural Network, Random Forest, Gradient Boosting, CatBoost, LightGBM, XGBoost, Balanced Bagging, and NGBoost. The performance evaluation of predictive models is done by employing fivefold cross-validation.

The MICE imputation technique shows slightly better performance compared to two alternative methods. LR-AGD excels on imbalanced data with the highest accuracy of 96.46% and XGBoost performs well on balanced datasets with the highest accuracy of 96.14%. However, their effectiveness is limited by dataset characteristics. In contrast, Random Forest delivers consistent and generalizable results with an accuracy rate around 95% on all non-augmented datasets. This characteristic becomes particularly evident when using the augmented dataset, it gives highest accuracy above 97%. After thorough evaluation, a more robust Dense Stacking Ensemble (DSE) model is constructed. The Random Forest model acts as the meta-classifier within the DSE model, with other models serving as base models post fine-tuning. The DSE model exhibits robust performance across both imbalanced and balanced MICE-imputed datasets. In the case of imbalanced dataset, the model achieved an accuracy of 96.13%, precision of 93.26%, recall of 96.18%, and an F1 score of 94.88%. Similarly, on balanced dataset, the DSE model achieved an accuracy of 96.59%, precision of 95.25%, recall of 96.27%, and F1 score of 95.79%. In terms of AUC, the DSE model achieved an AUC of 83.94% on the imbalanced dataset. On the balanced dataset, the DSE model excelled further by reaching an impressive AUC of 98.92%. These AUC scores demonstrates the ability of DSE model to distinguish between positive and negative instances. In conclusion, the DSE model consistently delivers robust and stable results for stroke prediction across diverse datasets.

In the future, the validation scope can be expanded with larger datasets that will encompass more features. Additionally, it is intended to explore diverse data formats which will include images and hybrid datasets. Furthermore, more extensive and diverse datasets will provide valuable insights and facilitate the generalizability of the findings to a broader population. By conducting external validation studies on these diverse and independent datasets, the authors aim to evaluate and validate the performance of the proposed DSE model. The proposed DSE model can integrate seamlessly into daily life using mobile or web applications, allowing the users to input health data effortlessly. With swift processing and prediction time the results can be relayed promptly. This integration will help in making prompt healthcare intervention in cases of predicted stroke risk.

Data availability

The McKinsey & Company's stroke prediction dataset for healthcare analytics is used in this study. The dataset is available publicly at Analytics Vidhya and Kaggle website at: https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon or https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Received: 26 February 2024; Accepted: 8 May 2024
Published online: 20 May 2024

References

1. Bersano, A. & Gatti, L. Pathophysiology and treatment of stroke: Present status and future perspectives. *Int. J. Mol. Sci.* **24**, 14848 (2023).
2. Feigin, V. L. *et al.* World stroke organization (wso): Global stroke fact sheet 2022. *Int. J. Stroke* **17**, 18–29 (2022).
3. Feigin, V. L. *et al.* Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the global burden of disease study 2019. *Lancet Neurol.* **20**, 795–820 (2021).
4. Katan, M. & Luft, A. Global burden of stroke. *Semin. Neurol.* **38**, 208–211 (2018).
5. Pitchai, R. *et al.* An artificial intelligence-based bio-medical stroke prediction and analytical system using a machine learning approach. *Comput. Intell. Neurosci.* **2022**, 1–9 (2022).
6. Amann, J. Machine learning in stroke medicine: Opportunities and challenges for risk prediction and prevention. *Adv. Neuroethics* https://doi.org/10.1007/978-3-030-74188-4_5 (2021).

7. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H. & Raad, A. Reviewing multimodal machine learning and its use in cardiovascular diseases detection. *Electronics* **12**, 1558 (2023).
8. Javaid, M., Haleem, A., Pratap Singh, R., Suman, R. & Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. *Int. J. Intell. Netw.* **3**, 58–73 (2022).
9. MacEachern, S. J. & Forkert, N. D. Machine learning for precision medicine. *Genome* **64**, 416–425 (2021).
10. Bonkhoff, A. K. & Grefkes, C. Precision medicine in stroke: Towards personalized outcome predictions using artificial intelligence. *Brain* **145**, 457–475 (2021).
11. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160 (2021).
12. Yu, J. *et al.* AI-based stroke disease prediction system using real-time electromyography signals. *Appl. Sci.* **10**, 6791 (2020).
13. Nijman, S. *et al.* Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *J. Clin. Epidemiol.* **142**, 218–229 (2022).
14. Kumar, Y., Koul, A., Singla, R. & Ijaz, M. F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Hum. Comput.* **14**, 8459–8486 (2022).
15. Kokkotis, C. *et al.* An explainable machine learning pipeline for stroke prediction on imbalanced data. *Diagnostics* **12**, 2392 (2022).
16. Sirsat, M. S., Fermé, E. & Câmara, J. Machine learning for brain stroke: A review. *J. Stroke Cerebrovasc. Dis.* **29**, 105162 (2020).
17. Wongvorachan, T., He, S. & Bulut, O. A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information* **14**, 54 (2023).
18. Sowjanya, A. M. & Mrudula, O. Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms. *Appl. Nanosci.* **13**, 1829–1840 (2022).
19. Bernat, J. L. & Lukovits, T. G. Ethical issues in stroke management. *Neurol. Clin. Pract.* **11**, 3–5 (2021).
20. Murdoch, B. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med. Ethics* https://doi.org/10.1186/s12910-021-00687-3 (2021).
21. Martin, C. *et al.* The ethical considerations including inclusion and biases, data protection, and proper implementation among AI in radiology and potential implications. *Intell. Based Med.* **6**, 100073 (2022).
22. Wu, Y. & Fang, Y. Stroke prediction with machine learning methods among older Chinese. *Int. J. Environ. Res. Public Health* **17**, 1828 (2020).
23. Kaur, M., Sakhare, S. R., Wanjale, K. & Akter, F. Early stroke prediction methods for prevention of strokes. *Behav. Neurol.* **2022**, 1–9 (2022).
24. Alanazi, E. M., Abdou, A. & Luo, J. Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models. *JMIR Format. Res.* **5**, e23440 (2021).
25. Monteiro, M. *et al.* Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 1953–1959 (2018).
26. Shobayo, O., Zachariah, O., Odusami, M. O. & Ogunleye, B. Prediction of stroke disease with demographic and behavioural data using random forest algorithm. *Analytics* **2**, 604–617 (2023).
27. Mezher, M. A. Genetic folding (GF) algorithm with minimal kernel operators to predict stroke patients. *Appl. Artif. Intell.* https://doi.org/10.1080/08839514.2022.2151179 (2022).
28. Guhdar, M., Ismail Melhum, A. & Luqman Ibrahim, A. Optimizing accuracy of stroke prediction using logistic regression. *J. Technol. Inform. (JoTI)* **4**, 41–47 (2023).
29. Sailasya, G. & Kumari, G. L. Analyzing the performance of stroke prediction using ml classification algorithms. *Int. J. Adv. Comput. Sci. Appl.* https://doi.org/10.14569/IJACSA.2021.0120662 (2021).
30. Paul, D., Gain, G. & Orang, S. Advanced random forest ensemble for stroke prediction. *Int. J. Adv. Res. Comput. Commun. Eng.* https://doi.org/10.17148/IJARCCE.2022.11343 (2022).
31. Geethanjali, T. M., Divyashree, M. D., Monisha, S. K. & Sahana, M. K. Stroke prediction using machine learning. *Int. J. Emerg. Technol. Innov. Res.* **8**, 710–717 (2021).
32. Harshitha, K. V., Harshitha, P., Gupta, G., Vaishak, P. & Prajna, K. B. Stroke prediction using machine learning algorithms. *Int. J. Innov. Res. Eng. Manag.* https://doi.org/10.21276/ijirem.2021.8.4.2 (2021).
33. Tazin, T. *et al.* Stroke disease detection and prediction using robust learning approaches. *J. Healthc. Eng.* **2021**, 1–12 (2021).
34. Chen, Z. Stroke risk prediction based on machine learning algorithms. *Highlights Sci. Eng. Technol.* **38**, 932–941 (2023).
35. Alruily, M., El-Ghany, S. A., Mostafa, A. M., Ezz, M. & El-Aziz, A. A. A-tuning ensemble machine learning technique for cerebral stroke prediction. *Appl. Sci.* **13**, 5047 (2023).
36. Islam, Md. M. *et al.* Stroke prediction analysis using machine learning classifiers and feature technique. *Int. J. Electron. Commun. Syst.* **1**, 57–62 (2021).
37. Uma, S. K. & Rakshith, S. R. Stroke analysis using 10 ml comparison. *Int. J. Res. Appl. Sci. Eng. Technol.* **10**, 3857–3862 (2022).
38. Fedesoriano. Stroke prediction dataset. *Kaggle*. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data (2021).
39. Mattas, P. S. Brain stroke prediction using machine learning. *Int. J. Res. Publ. Rev.* **3**, 711–722 (2022).
40. Pathan, M. S., Jianbiao, Z., John, D., Nag, A. & Dev, S. Identifying stroke indicators using rough sets. *IEEE Access* **8**, 210318–210327 (2020).
41. Emon, M. U. et al. Performance Analysis of Machine Learning Approaches in stroke prediction. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (2020).
42. Hassan, A. & Yousaf, N. Bankruptcy prediction using diverse machine learning algorithms. In *2022 International Conference on Frontiers of Information Technology (FIT)* (2022).
43. Dev, S. *et al.* A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthc. Anal.* **2**, 100032 (2022).
44. Nguyen, L. P. *et al.* The utilization of machine learning algorithms for assisting physicians in the diagnosis of diabetes. *Diagnostics* **13**, 2087 (2023).
45. Mahajan, P., Uddin, S., Hajati, F. & Moni, M. A. Ensemble learning for disease prediction: A review. *Healthcare* **11**, 1808 (2023).
46. Jagan, S. *et al.* A meta-classification model for optimized zbot malware prediction using learning algorithms. *Mathematics* **11**, 2840 (2023).
47. Zhen, M. *et al.* Application of a fusion model based on machine learning in visibility prediction. *Remote Sens.* **15**, 1450 (2023).
48. Yuan, Q., Chen, K., Yu, Y., Le, N. Q. & Chua, M. C. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac630 (2023).
49. Le, N.-Q.-K. & Ou, Y.-Y. Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins. *BMC Bioinform.* https://doi.org/10.1186/s12859-016-1369-y (2016).

Acknowledgements

The authors express gratitude to the anonymous reviewers for their insightful comments and suggestions, contributing to the enhancement of this paper.

Author contributions

Ahmad Hassan wrote the main manuscript text, Saima Gulzar Ahmad and Ehsan Ullah Munir developed figures, and Naeem Ramzan contributed to the results section. Imtiaz Ali Khan conducted additional simulations and included supplementary results. All authors reviewed and enhanced the manuscript write-up. The final manuscript was read and approved by all authors.

Funding

This research was partially supported by the SAFE-RH project under Grant No. ERASMUS + CBHE-619483 EPP-1-2020-1-UK-EPPKA2-CBHE.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.R.

Reprints and permissions information is available at www.nature.com/reprints.

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