World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Adopting random forest for predicting the risk of cerebrovascular disease and diabetes using appropriate database

Kingsley Kwesi Acheampong [1, *] and Zhou Jinzhi [2]

[1] *College of Engineering, Northeastern University Boston, Massachusetts, USA.*

[2] *Department of Information and Communication Engineering, Southwest University of Science and Technology, Mianyang, Sichuan, China.*

World Journal of Advanced Research and Reviews, 2024, 24(03), 2888-2899

Publication history: Received on 16 November 2024; revised on 26 December 2024; accepted on 28 December 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.3.3999

Abstract

Random forest is used to predict the possibility of the existence of a cerebrovascular disease subjects curated from the BraVa and diabetes datasets. Towards analyzing and prediction of cerebrovascular diseases, SPSS for realizing the independence and correlation between the various metrics that could contribute to a subject been diagnosed with a cerebrovascular disease. An analysis of the various metrics of the overall vascular size revealed a significant correlation especially between Total Length and Total Number of Branches (R = 0.829, p = 0.000). Metrics like Age, Contraction, Tortuosity, mean bifurcation Angle, mean bifurcation tilt which has implication of a cerebrovascular disease diagnosis according to study was used as the input for the random forest algorithm. The BraVa dataset which is the main datasets for this work was used to train the algorithm and a prediction of either "risky" or "Not risky" with a high accuracy of 100% was recorded. To further test the algorithm, a second datasets from the from the diabetes database which has a high number of subjects was also used to test the algorithm and a high accuracy of 90.256% was recorded. It was determined from the results that machine learning based Random Forest algorithm can be adopted as a prediction method especially on bigger dataset of neuromorphological measurements of neurons and it will aid or facilitate accurate prediction of any form of cerebrovascular disease and also aid in accurate medical diagnosis.

Keywords: Random Forest; Cerebrovascular disease; Diabetes; Machine Learning; Brain Vasculature

1 Introduction

Cerebrovascular diseases (CVD) are the leading cause of death in humans worldwide, crippling morbidity and long-term disability. Variations in the neurovascular structure–function relationship between individuals and groups have not yet been fully investigated. Quantitative characterization of cerebrovascular architecture from modern magnetic resonance angiography (MRA) may lead to a better understanding of the cerebrovascular system's physiological role and pathological dysfunction. MRA is a non-invasive procedure for the visualization of cerebral arteries in three dimensions. The distinction between fast-moving arterial blood and stagnant tissues that surround the artery is based on this. To date, most MRA studies have been limited to qualitative or semi-quantitative evaluations, partial morphometric analyzes small numbers of subjects, and proprietary datasets (1). Reconstruction of vascular arborization into an explicit 3D representation will achieve a more detailed structural characterization of the cerebral arterial tree (2, 3). In addition to allowing comprehensive morphometric research, these reconstructions can be used with fluid dynamics modeling for subject-specific assessment of the individual risks of vascular malformation (4, 5). Such methods include specification of correct boundary conditions and constraints relevant to geometry of arterial branches and characteristics of bifurcation (6). The complexity of manually reconstructing a broad vascular network, however, limited numerical simulations to synthetic arterial tree models (7).

[*] Corresponding author: Kingsley Kwesi Acheampong

Human brains change with aging. Changes include progressive shrinkage of gray matter volume (8), changes in white matter such as fractional anisotropy (9) and the loss of myelinated axons (10). These changes could be caused by changes in the brain vasculature, such as plaque formation and microhemorrhage, microvascular disease (11) and age-related decline in capillary number (12).

For decades, Random Forest method is applied in various fields. Random forest (RF) is an extension of the bagging method a learning system typical of the ensemble (13). The method typically picks a sample at random and places it into the sample collection, and the sample is then put back into the original data collection, so that the sample can still be picked at the next sampling time. In this way we get a sample collection of m samples after m random sampling operations. Some samples appear several times in the resampling collection as part of the initial training package, and some never appear. T samples containing m samples of training are chosen, a basic learner is then trained based on each sample collection, and these basic learners are then combined. RF's base learner is a decision tree, and random selection of attributes are incorporated into the decision tree training process.

RF is simple, comprehensible, computationally inexpensive, and has achieved powerful output in many real-world tasks. Montesinos et al (14), reported on the application of RF for genomic prediction using data from plant breeding. In a similar study, Osval et al (15) reported on a comparative study of conventional random forest and an improved model (zero altered poison random forest) for gene prediction. Pazhanikumar and KuzhalVoiMozhi also adopted the RF model to classify remotely sensed images utilizing three datasets; SAT-4, SAT-6 and RSI-CB (16). In the study, images are categorized using a majority voting procedure on a tree-based structure using a modified Random Forest (RF) with an empirical loss function. Loss values are calculated to assess the model's effectiveness.

In this work, a statistical approach is firstly used to determine by pairwise correlations with Pearson’s coefficient, with the p values indicating the probability of independent distributions. The correlation between age and some metrics of the vasculature which could contribute to cerebrovascular disorders was realized. Further, a demonstration of how the statistical analysis method can be used to study individual and population differences in cerebral vasculature at the level of the entire vasculature, specific arteries or single branches. Age-related changes, hemispheric lateralization, and gender-related difference in cerebral circulation may all be important risk factors in cerebrovascular disorders (17). For instance, increase in tortuosity of right arterial trees during normal aging may have clinical implications (18). Secondly a machine learning based random forest decision tree algorithm was further used to prove and predict based on these numerical values the effects or likelihood of cerebrovascular disorder. The methodology is used to study and predict the effect of the various metrics on brain vasculature in a collection of healthy 44 subjects. These techniques could also be used for other studies of vascular morphology and their effect on clinical outcomes

2 Material and methods

2.1 Proposed Method

The methodology employed in this research integrates a series of systematic processes to ensure the development of an accurate and reliable model for analysis and prediction. As depicted in Figure 1, the framework encompasses data acquisition, preprocessing, feature selection, model training, testing, and evaluation. The implementation of these steps provides a structured approach to enhancing the predictive capacity of the Random Forest (RF) algorithm utilized in this study.

2.1.1 Data Loading and Description

The study utilized two distinct datasets: the Brain Vasculature (BraVa) dataset and the Diabetes Classification dataset. These datasets served as the foundation for model training and testing, ensuring a robust evaluation of the proposed approach.

2.1.1.1 Brain Vasculature Dataset

The BraVa dataset, obtained from http://cng.gmu.edu/brava, contains digitally reconstructed human arterial arborizations derived from the circle of Willis. This dataset includes six primary arterial branches: the left and right Anterior Cerebral Arteries (ACAs), Middle Cerebral Arteries (MCAs), and Posterior Cerebral Arteries (PCAs). For this study, a subset of the dataset comprising 44 healthy adult subjects aged 19–59 years was selected. Individuals with conditions such as diabetes, hypertension, head trauma, psychiatric disorders, or other health issues that could influence brain vasculature were excluded to maintain dataset integrity.

2.1.1.2 Diabetes Classification Dataset

The second dataset was derived from the Vanderbilt Biostatistics program and accessed through Data.World (http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets). This dataset consists of real-world patient data tailored to facilitate the classification of diabetes (binary classification: “yes” or “no”) based on demographic and laboratory variables. The data underwent preprocessing, where patients with hemoglobin A1c values less than 6.5% were excluded, and those with values equal to or greater than 6.5% were labeled as diabetic. Out of 390 records, 60 samples were identified as diabetic and included for further analysis.

Both datasets were converted into CSV file formats for seamless integration into the model development pipeline.

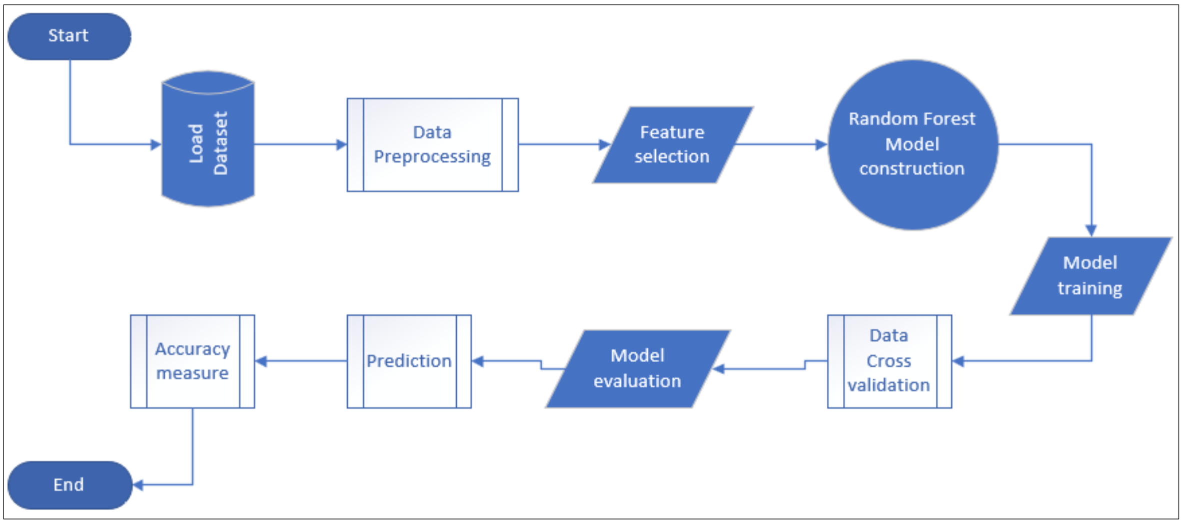


Figure 1 Process diagram showing the various steps involved in the random forest implementation

2.1.2 Data Preprocessing

To ensure the datasets were clean, consistent, and suitable for analysis, preprocessing was conducted. This step involved handling missing data, normalizing features, and transforming variables where necessary. These measures were undertaken to eliminate noise, reduce (19) redundancy, and enhance the reliability of the predictive model.

2.1.3 Feature selection

Feature selection was performed to identify the most relevant attributes for model training. This process aimed to reduce dimensionality, improve computational efficiency, and minimize overfitting. The interdependencies between features were evaluated to ensure that only the most significant predictors were included in the modeling phase. (20)

2.1.4 Model Construction

The construction of the Random Forest (RF) model is predicted on the assumption that the training set contains samples and the total number of variables in the dataset is. A critical aspect of this step involves selecting a subset of input variables to determine decisions at each tree node. The subset size (where) is chosen to be significantly smaller than, thereby enhancing the diversity and accuracy of the resulting decision trees.

The training process initiates by selecting samples from the training dataset with replacement, creating multiple bootstrap samples. For each node of the decision tree, variables are randomly selected from the total variables. These variables are utilized to compute the optimal split point at each node. This iterative process continues until the tree is fully grown, ensuring that the decision trees are not pruned. This approach allows RF to generate a diverse ensemble of decision trees by leveraging randomization during both sample selection and node splitting.

Each tree in the ensemble is trained using approximately two-thirds () of the original training data, a process known as bagging or bootstrap aggregation. The remaining one-third (1/3) of the data, termed “out-of-bag” (OOB) data, is reserved for validating the performance of the individual trees within the ensemble. This technique provides an unbiased estimate of the test set error, enabling an accurate assessment of the RF model.

Additionally, RF introduces another level of randomization during the splitting of decision nodes. Unlike traditional tree algorithms, such as Classification and Regression Trees (CART), which consider all available variables to determine splits, RF restricts the search to only randomly selected variables at each split. This ensures that the ensemble comprises trees with a high degree of variability, mitigating overfitting and improving generalization.

The combined effects of bagging and randomized node splitting are fundamental to the RF algorithm's performance. By aggregating predictions from multiple independent trees, RF minimizes variance and improves predictive accuracy. The dual sources of randomness bootstrap sampling and variable selection are key attributes that distinguish RF as a robust and efficient machine learning model.

This process is illustrated in Figure 2, which depicts the workflow of the RF model construction, demonstrating how multiple decision trees are independently trained and aggregated to produce the final predictive output.

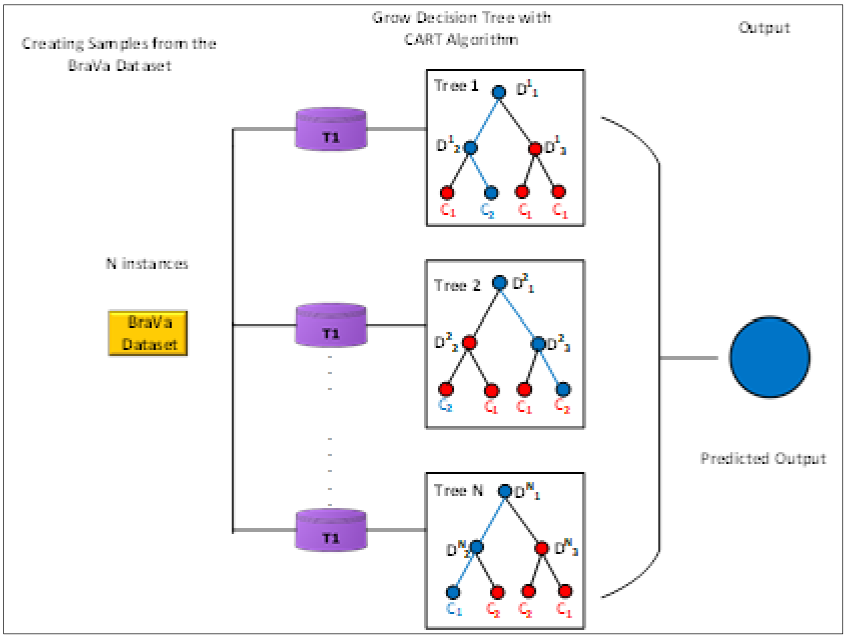

Creating Samples from the BraVa Dataset
Grow Decision Tree with CART Algorithm
Output
N instances
BraVa Dataset
T1
T1
T1
Tree 1
Tree 2
Tree N
Predicted Output

Figure 2 Random Forest algorithm Model design

2.1.5 Model Training and Testing

The training process of the Random Forest (RF) model in this study is conducted using the BraVa dataset to enable the algorithm to accurately learn the inherent patterns within the data. The training phase involves fitting the RF algorithm to the training set, ensuring that it identifies relationships and dependencies within the data effectively.

One of the key considerations during model training is the selection of parameters that significantly influence the model's performance. These parameters include the number of decision trees in the RF and the splitting criterion used to construct each tree. The number of trees chosen for this study ranges from 5 to 80. While there is no fixed rule for the optimal number of trees, careful consideration is required to strike a balance between mitigating overfitting and underfitting.

The splitting criterion used in this study is the Gini Index, a measure of impurity that determines the attribute to be used for splitting data at each node. The Gini Index is calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \qquad (2)$$

Where P_i is the probability that a tuple in D belongs to class Ci and is estimated by $|C_{iD}|/|D|$.The sum is computed over m classes. The attribute that reduces the impurity to the maximum level (or has the minimum Gini_ index) is selected as the splitting attribute.

2.1.6 Data Cross Validation

To optimize the hyperparameters of the RF model and evaluate its generalization performance, a cross-validation (CV) technique is employed. The RF hyperparameters include the number of trees in the ensemble and the subset of features considered at each split. Proper tuning of these hyperparameters is critical to avoid overfitting, where the model performs exceptionally well on the training set but poorly on unseen data.

This study adopts the n-Fold Cross-Validation method, which divides the training dataset into subsets, or "folds." Each fold acts as a validation set while the remaining folds are used for training. This process is repeated times, ensuring that each subset is used as validation exactly once. For instance, in 5-Fold Cross-Validation, the data is split into five equal parts:

- The first iteration uses four folds for training and the fifth for validation.
- The second iteration trains on folds one, two, three, and five, while the fourth serves as validation.
- This process repeats until all folds are used for validation.

The validation results from all folds are averaged to provide an unbiased estimate of the model's performance.

In this study, 5-Fold Cross-Validation is employed to assess the accuracy and robustness of the RF model. By evaluating the model across multiple folds, the technique ensures that the results are reliable and generalizable to unseen data. This methodology enables the study to estimate the model's predictive capability effectively, thereby validating its overall performance. The cross-validation framework and results are further illustrated in Figure 3.

In the case of a random forest, the hyperparameters include the number of decision trees in the forest and the number of characteristics that each tree considers when dividing a node. Deciding the best hyperparameters in advance is usually difficult, and tuning a model is where machine learning turns from a science into trial-and-error based engineering. Evaluating the model only on the training set would give rise to one of the most important issues of machine learning overfitting. An overfit model can look amazing on the training set but in a real-world application it will be useless. Hence the basic technique for optimizing the hyperparameter accounts for overfitting through cross validation. Cross-validation (CV) technique is best described by using the most common form, n-Fold CV. There is a split of the training in n-Fold CV set up into n number of subsets, called folds. This is then iteratively applied to the model n times, each time the fold data are trained on n-1 and tested on the nth fold (called the validation data). For example, the first iteration trains on the first four folds in fitting a model with n = 5, and evaluate on the fifth. The second iteration trains and tests on the fourth, on the first, second, third and fifth layer. The average of the results on each fold is assessed at the very end of the training to arrive at the final validity metrics for the model.

In this study, the 5-fold cross validation is adapted to estimate the performance of the learned model relative to unseen results (Figure 3). This helps in estimating the model's results.

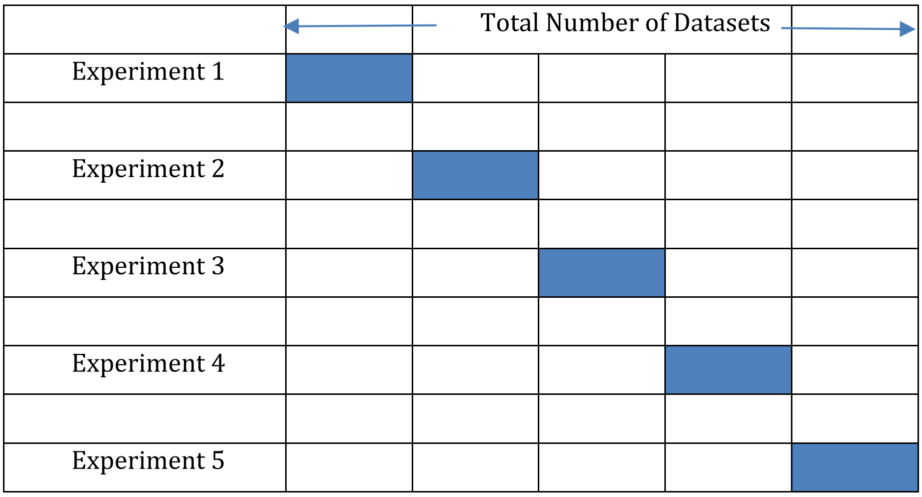


Figure 3 5-Fold Cross Validation.

2.1.7 *Model Evaluation*

As mentioned above, a smaller number of datasets was used to train the model. Therefore, in order to really evaluate the model to see how best its performance is, the bigger dataset which has a large number of features than the training dataset was used to test the algorithm. The model looks at the test data point and then learns a little more about the relationships between the features and labels. Assuming that there are relationships in the data giving the model more data will allow it to better understand how to map a set of features to a label. During model evaluation, the performance of the learned model is assessed using such techniques as prediction and accuracy measurement.

2.1.7.1 Prediction

Prediction here is performed using the trained random forest algorithm which passes the test features through the rules of each randomly created trees. Typically, with this model where 80 random trees were selected, each of the random forest will predict different targets (outcomes) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 80 random decision trees predicts some 3 unique targets **x, y, z** then the votes of x are nothing but out of 80 random decision trees how many trees prediction is **x.**

This is the same for (y and z) if the number of votes of x is higher. Let's say out of 80 random decision tree **60** trees are predicting the target will be x. Then the final random forest returns the x as the predicted target. This concept of voting is known as **majority voting**. At this stage the test data from the original dataset are passed to the learned model so as to make prediction from the decision trees. At test time, predictions are made by averaging the predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as bagging, short for bootstrap aggregating.

2.1.7.2 Accuracy Metric

Based on the actual and predicted values, a measure of the accuracy is computed by finding percentage of correct predictions as against the actual values from the dataset.

$$Accuracy = \frac{TP+NP}{TP+TN+FP+FN} \quad (3)$$

Where, True positive (TP), means the test predicts that the patients have CVD, and they have CVD. False positive (FP) means the test predicts that the patients do not have CVD, but they do. True negative (TN) implies the test predicts that the patients do not have CVD, and they do not have CVD. False negative (FN,) the test predicts that the patients have CVD, but they do not.

3 Results and discussion

Statistical Analysis of the data were done using IDM Statistical Package for Social Sciences (IBM SPSS v20). Specifically, the means and standard deviations were calculated using the Frequencies tool. Correlation coefficient and significant differences of the data for the various parameters were determined by the Pearson model. Microsoft Excel was used in the plotting of statistical graphs for data result presentation. A Lenovo ThinkPad with processor Intel(R) Core(TM) i7-5600U CPU @2.60GHz 2.59GHz and RAM of 8.00GB which is a 64-bit operating system was used for this work.

3.1 Quantitative Anatomy of Cerebral Arteries

A summary sstatistics for the various scaler parameters that characterize the entire vascular structure with regards to overall size, branch features and bifurcation angles and symmetry are computed and displayed in Table 1. Overall size variability of the data used in this study was similar to the values reported for other parameters of human body size. For instance, the coefficients of variation for total number of branches and total length were approximately between 0.13 and 0.25, and associated ranges between 68% - 157% of the means of the respective parameters. Analysis of the various metrics of the overall vascular size revealed a significant correlation especially between Total Length and Total Number of Branches ($R = 0.829$, $p = 0.000$). A statistically significant difference also existed between Total Length and MBO ($R = 0.323$, $p<0.035$), Height ($R = 0.354$, $p<0.02$), Depth ($R = 0.438$, $p<0.003$) and Tortuosity ($R = 1.000$, $p=0.000$). MBO also had a significant difference with PA ($R = 0.323$, $p<0.035$), BPL ($R = -0.308$, $p<0.045$) and Tortuosity ($R = 0.323$, $p<0.032$). Studies have shown that these parameters are critical in the determination of a person's risk of getting cerebrovascular diseases.

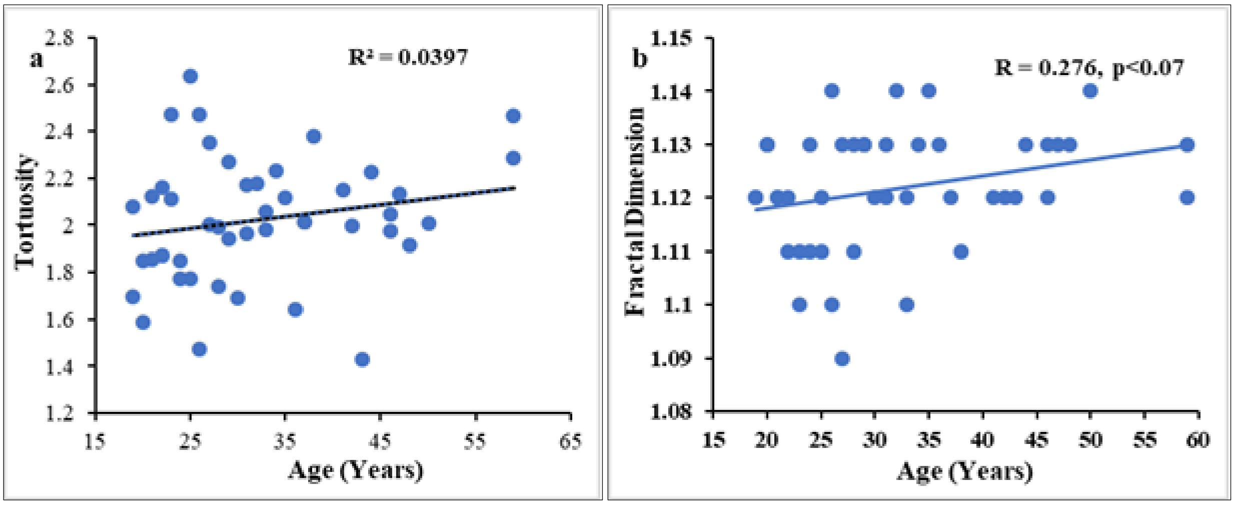
Aneurysms is highly prevalent at or close to bifurcations, hence making bifurcation an important area for focus in the determination of vascular diseases. A substantially lower variability is displayed in the sample means of averages within arbors. The coefficient of variation for mean bifurcation amplitude is lower than 0.06 and its values range from 87.4% to 113.2%. The angular value (89.9°) for mean local bifurcation amplitude was seen to be approximately 1.5 times greater than the mean bifurcation amplitude value (61.2°). An analysis of the correlation by Pearson model showed an insignificant negative correlation between the two parameters.

Analysis of the data showed a significant correlation which studies have revealed to have significant on the risk of a person getting cerebrovascular diseases such as stroke. A negative and significant correlation was found to exist between age and contraction (R = -0.424, p = 0.004) as well as age and mean bifurcation tilt (MBT) (R = -0.506, p = 0.000). The significant correlation between age and the parameters (MBA, MBT and Contraction) gives an indication of the risk of a person getting a cardiovascular disease with age. Tzimas et al (21) reported on the increase in the risk of CVD with age due to related arterial plaque. This cause high shear stresses being exerted on vessel walls by blood, thereby increasing the risk of stroke. Also, as a function of age, a significant positive correlation was seen between mean bifurcation angle (MBA) (R = 0.435, p = 0.003). This means that MBA increases with an increase in age.

Table 1 Whole Arterial Metric

Item	Metric	Overall (N=44) μ ± σ (min – max)
Overall size	Total Number of Branches (TNB)	211.1 ± 41.0 (144 – 330)
	Total Length (TL) (mm)	7050.4 ± 934.6 (4978.1 – 9171.0)
	Max Branch Order (MBO)	15.5 ± 1.3 (13 – 18)
	Max Path Distance (MPD) (mm)	285.9 ± 20.7 (250.0 – 348.4)
	Max Euclidean Distance (MED) (mm)	109.1 ± 4.8 (97.7 – 118.8)
	Width (mm)	116.5 ± 17.1 (14.1 – 147.2)
	Height (mm)	84.8 ± 3.7 (75.6 – 93.3)
	Depth (mm)	129.6 ± 10.6 (93.3 – 145.7)
Bifurcation Amplitude	Mean Bifurcation Angle (MBA) (°)	61.2 ± 3.3 (53.5 – 69.3)
	Mean Local Bifurcation Angle (MLBA) (°)	89.4 ± 3.3 (83.2 – 98.6)
	Mean Bifurcation Tilt (MBT)	103.8 ± 4.4 (92.4 – 113.3)
Branch	Partition Asymmetry (PA)	0.5 ± 0.1 (0.4 – 0.6)
	Branch Path Length (BPL)	34.2 ± 3.5 (27.6 – 43.2)
	Mean Fragmentation (MF)	15.4 ± 1.6 (12.3 – 19.5)
	Fractal Dimension (FD)	1.1 ± 0.0 (1.1 – 1.1)

Another feature critical in the determination of cerebrovascular diseases is the branch. The tortuosity of a branch gives an indication of the type of cerebrovascular diseases such as stroke, diabetes and hypertension. A positive correlation was observed between age and tortuosity, fractal dimension and path length (Figure 1 a&b). In contrast, a negative correlation was observed between age and path asymmetry. No statistically significant difference was observed for the other parameters except between branch path length and fractal dimension (34.15 ± 3.52 and 1.12 ± 0.01).



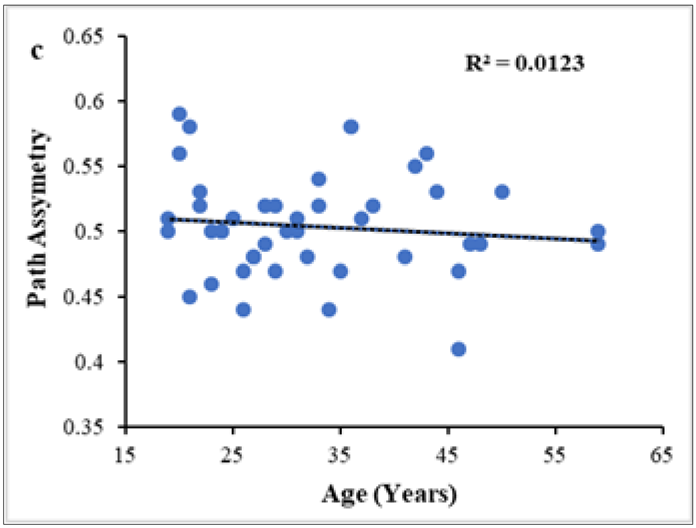


Figure 4 Scatter plot showing the correlation between age and morphometric features (a) Tortuosity (b)Fractal dimension and (c) Path asymmetry

3.2 Size Differences and Proportional Scaling among Cerebral Arteries

The scalar morphometric characteristics of singular cerebral arteries is highlighted in Table 2. The analysis of the obtained data indicated that values increased in the order of PCA → ACA → MCA. For instance, the total number of branches for MCA was twice thrice as much as that of ACA and PCA respectively. However, the width, height and depth does not follow the same systematic pattern due to the differences in the orientations of cerebral arteries in relation to the brain's canonical planes.

Conducting in-depth analysis is critical in the characterizing arterial architecture in individual brains. A significant correlation was identified between the left and right sides of the various parameters. For instance, the R and p values of for the left and right sides of TL for MCA, ACA and PCA are R = 0.677 and 0.000, 0.485 and 0.003, and 0.729 and 0.00 respectively. Figure 2a, displays the linear relation existing between the left and right lengths of the MCA, ACA and PCA.

The composition of arterial length within-subject is examined to determine if compensation is made for cerebral arteries with regards to the total length (Figure 2b). A positive significant correlation was observed to exist among the length each artery and the total vascular arborization (the R value was 0.861, 0.452, and 0.618 for MCA, ACA and PCA respectively. $p < 0.007$). This result indicates that there is a proportional scaling of arteries in the brain. That is, if a big vasculature exists, all the arteries are big.

The neuronal branching contributes a significant effect of the risk of cerebrovascular disease. The analysis showed a negative correlation between Age and Contraction (R=-0.424, p =0.004). It's been proposed that older ages have high tendencies for arterial plaque resulting in a narrow artery diameter. Hence the morphological measurements of these metrics were curated ass Age/MBT, Age/Contraction and Age/Tortuosity and then labelled according the mean of the various values to distinguish between what the algorithm sees as risky or not risky and then run through the random forest algorithm.

Table 2 Individual Artery Metrics

Item	**Matrix**	**Overall (N=44) μ ± σ (min – max)**		
		PCA	ACA	MCA
Overall Size	TNB	37.9 ± 11.9 (14-64)	50.7 ± 12.2 (16-74)	101.5 ± 19.9 (34-144)
	TL (mm)	1015.1 ± 222.1 (510.6 – 1481.1)	1760.1 ± 309.8 (1262.5 – 2547.2)	3716.5 ± 510.5 (2289.7–4699.4)
	MBO	11.7 ± 3.1 (5 – 10)	12.3 ± 2.0 (9 – 18)	17.5 ± 1.7 (13 – 21)
	MPD	300.6 ± 29.0 (248.1 – 371.2)	371.6 ± 30.7 (304.1 – 438.0)	448.4 ± 37.6 (381.2 – 536.6)
	MED	180.9 ± 13.5 (151.0-201.3)	191.8 ± 14.3 (163.0 – 218.6)	197.8 ± 10.6 (168.9 – 216.3)

	Width (mm)	63.1 ± 11.7 (45.6 – 88.7)	51.6 ± 7.0 (40.9 – 72.2)	77.7 ± 9.6 (67.3 – 127.1)
	Height (mm)	102.1 ± 22.5 (43.4 – 144.2)	142.0 ± 21.9 (51.8– 190.3)	136.5 ± 14.8 (71.0 – 155.6)
	Depth (mm)	130.3 ± 17.6 (58.6 - 158.1)	217.8 ± 23.8 (149.4 -269.1)	209.6 ± 15.4 (177.9 -236.2)
Bifurcation Amplitude	MLBA (°)	187.2 ± 21.5 (127.7 – 226.1)	181.3 ± 16.1 (140.4 – 212.6)	174.0 ± 10.1 (156.7 – 191.5)
	MBA (°)	117.6 ± 18.8 (91.4 - 183.1)	103.3 ± 14.1 (75.3- 152.4)	113.7 ± 14.3 (91.4-172.7)
	MB Tilt (°)	233.6 ± 20.2 (183.3 – 267.4)	215.6 ± 33.0 (95.9 – 259.2)	202.6 ± 12.4 (177.9 – 229.2)
	MB Torque (°)	172.8 ± 24.0 (136.2 – 227.8)	164.8 ± 25.4 (95.9 – 259.2)	172.9 ± 12.4 (142.4 – 192.0)
Branch	PA	1.1 ± 0.2 (0.7 – 1.5)	1.0 ± 0.2 (0.6 – 1.4)	1.0 ± 0.1 (0.7 – 1.3)
	BPL	59.5 ± 11.0 (41.2 – 89.1)	74.1 ± 11.2 (49.2 – 102.6)	74.3 ± 7.8 (59.8 – 92.7)
	MF	29.2 ± 5.6 (19.6 – 46.5)	35.7 ± 5.4 (24.6 – 50.0)	32.2 ± 3.4 (25.8 – 40.1)
	FD	2.2 ± 0.0 (2.1 – 2.3)	2.1 ± 0.3 (0.9 – 2.3)	2.3 ± 0.0 (2.2 – 2.3)
	Contraction	1.4 ± 0.1 (1.3 – 1.7)	1.4 ± 0.1 (1.3 – 1.6)	1.3 ± 0.1 (1.2 – 1.4)

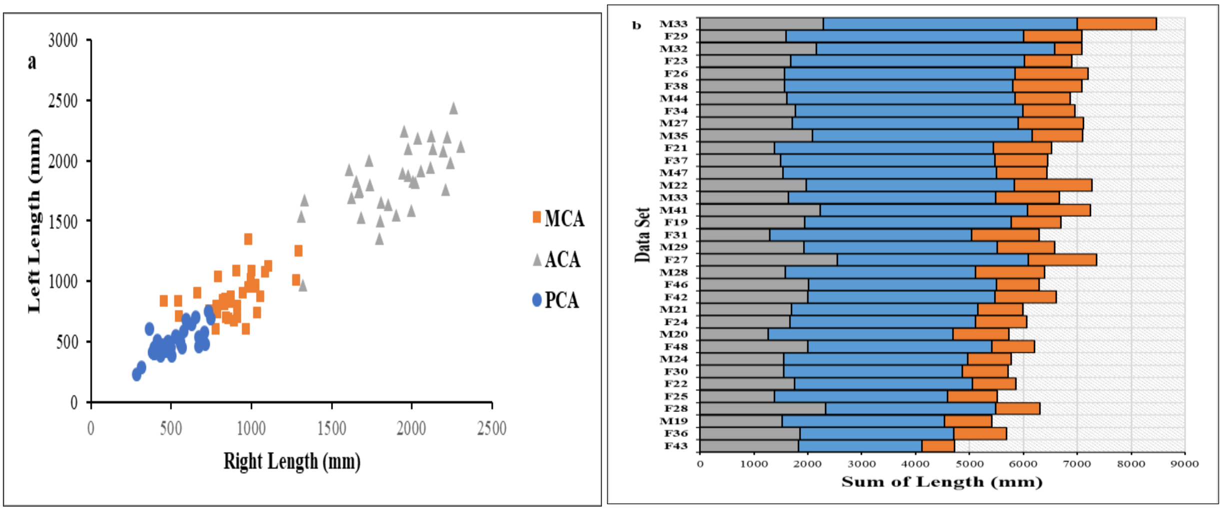


Figure 5 Graphical representation of (a) linear relation existing between the left and right lengths of the MCA, ACA and PCA; (b)composition of arterial length within-subject

3.3 Performance of Random Forest on Brava and Diabetes datasets

Random Forest is the final machine learning method that was used. It also contains a hyper-parameter that can be tuned to the data, which is the number of trees to have. The selection of trees is important, as it determines the sample of observations to use. Not having enough trees can mean that some observations of a class may not get selected at all in the training process. The following values for the number of trees: [5, 15, 25, 40, 65, 80] was used for optimization. Having too many trees significantly increases the computation time and power needed. Therefore, a maximum of 65 trees is selected for optimization.

3.3.1 Random forest on Brava dataset

The performance of each selection of trees, using Age/Contraction, Age/MBT, and Age/Tortuosity datasets curated from the BraVa database. Having 80 trees does give the best performance of 100% mean accuracy and predicted 8 out of the actual 8 value, 97.5% and predicted 8 out of the actual 8, also 92.5% and 8 prediction out of 8 actuals for the BraVa database. In general, the more trees you use the better the results. However, the improvement decreases as the number of trees increases. That is, in Age/Contraction it was realized that from tree 5 to 10, there is a decrease in the accuracy and then again as the tree grew from 25 to 40 the accuracy again increased. From the results, there is an indication that the higher number of trees do further improve the performance.

This explains that the algorithm was able to predict with the highest accuracy of 100% whether or not a particular subject is at risk or not of being affected by a cerebrovascular disease. Further explanation of the test or algorithm shows that, for Age/contraction, there was 43 total rows and columns. High Age and Low Contraction was labelled “Risky”. Then low Age and high contraction labelled “Not Risky”. After running the data, the algorithm was able to predict according to the number of trees with a given accuracy as shown above and for all the trees and final prediction was accurate compared to the actual. This was done for all the other datasets and the various mean accuracy values were recorded below in the table.

Table 3 Mean accuracy values of the various datasets from the algorithm

Number of trees	Age/Contraction (%)	Age/MBT (%)	Age/Tortuosity (%)
5	100	95	90
15	97.5	95	90
25	95	97	90
40	97.5	95	92.5
65	100	95	92.5
80	100	97.5	92.5

3.3.2 Testing Random Forest Algorithm with the Diabetes Dataset

The table above shows the performance of each selection of trees, using Diabetics datasets. Having 80 trees does give the best performance of 90.256% mean accuracy and predicted 8 out of the actual 8 value, 90.0% and predicted 8 out of the actual 8, then 90.0% and 8 prediction out of 8 actuals for and the least of 89.487% and predicted 8 out of 8 actuals. The results indicate that the higher number of trees do further improve the performance. The algorithm however was able to correctly make final prediction of the patient being diabetic or not with a good accuracy of 90.3% at the 65 trees. As stated earlier, for optimization purposes tree number 65 was chosen which gives a good prediction and accuracy from the dataset above.

Table 4 Mean accuracy values of the diabetes datasets from the algorithm

Number of trees	Mean Accuracy Values (%)
5	89.5
15	89.7
25	90.0
40	90.3
65	90.3
80	90.0

There are some differences in the mean accuracy values from the algorithm running both the BraVa dataset and the Diabetes dataset. One may ask why? The answer is that the algorithm has been designed to be dynamic to take every kind of data once the data has the same format is in a certain value which you want to predict the output. In spite of the same number of trees being 80, the accuracy values from the BraVa dataset were a bit higher than those from the diabetes database because the Brava dataset which is the primary data contains smaller entities than that of the Diabetes data.

4 Conclusion

The aim of this study is to develop an efficient model with improved accuracy to predict cerebrovascular disease and diabetes. Datasets relevant to the mentioned conditions were obtained and statistically analyzed using correlation coefficient. From the analysis, the correlation of age with contraction, MBT and tortuosity were revealed to significantly determine CVD. Performance evaluation of the proposed model based on the critical features showed a mean accuracy value of 100%, 97.5% and 92.5% respectively for Age/Contraction, Age/MBT and Age/Tortuosity at 80 trees. Meanwhile, a mean accuracy value of 90.3% for the diabetes dataset was achieved at 65 trees. The high mean accuracy values for the two datasets indicates the effectiveness and dynamic nature of the proposed algorithm.

Compliance with ethical standards

Disclosure of Conflict of interest

All authors declare there are no known conflict of interest

References

[1] Dong QL, Barker GC, Gorris LGM, Tian MS, Song XY, Malakar PK. Status and future of Quantitative Microbiological Risk Assessment in China. Trends in Food Science & Technology. 2015;42(1):70-80.

[2] Bullitt E, Muller KE, Jung I, Lin W, Aylward S. Analyzing attributes of vessel populations. Medical image analysis. 2005;9(1):39-49.

[3] Passat N, Ronse C, Baruthio J, Armspach JP, Maillot C. Magnetic resonance angiography: from anatomical knowledge modeling to vessel segmentation. Medical image analysis. 2006;10(2):259-74.

[4] Cebral JR, Castro MA, Soto O, Löhner R, Alperin N. Blood-flow models of the circle of Willis from magnetic resonance data. Journal of Engineering Mathematics. 2003;47(3):369-86.

[5] Oshima M, Torii R, Kobayashi T, Taniguchi N, Takagi K. Finite element simulation of blood flow in the cerebral artery. Computer Methods in Applied Mechanics and Engineering. 2001;191(6):661-71.

[6] Olufsen MS. Structured tree outflow condition for blood flow in larger systemic arteries. The American journal of physiology. 1999;276(1):H257-68.

[7] Bui AV, Manasseh R, Liffman K, Sutalo ID. Development of optimized vascular fractal tree models using level set distance function. Medical engineering & physics. 2010;32(7):790-4.

[8] Yang Z, Wen J, Erus G, Govindarajan ST, Melhem R, Mamourian E, et al. Brain aging patterns in a large and diverse cohort of 49,482 individuals. Nature Medicine. 2024;30(10):3015-26.

[9] Jáni M, Mareček R, Mareckova K. Development of white matter in young adulthood: The speed of brain aging and its relationship with changes in fractional anisotropy. NeuroImage. 2024;301:120881.

[10] de Faria O, Pivonkova H, Varga B, Timmler S, Evans KA, Káradóttir RT. Periods of synchronized myelin changes shape brain function and plasticity. Nature Neuroscience. 2021;24(11):1508-21.

[11] Fernando MS, Simpson JE, Matthews F, Brayne C, Lewis CE, Barber R, et al. White matter lesions in an unselected cohort of the elderly: molecular pathology suggests origin from chronic hypoperfusion injury. Stroke. 2006;37(6):1391-8.

[12] Allaf AM, Wang J, Simms AG, Jiang H. Age-related alterations in retinal capillary function. Microvascular research. 2023:104508.

[13] Mohapatra N, Shreya K, Chinmay A. Optimization of the Random Forest Algorithm. 2020. p. 201-8.

[14] Montesinos López OA, Montesinos López A, Crossa J. Random Forest for Genomic Prediction. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Cham: Springer International Publishing; 2022. p. 633-81.

[15] Montesinos-López OA, Montesinos-López A, Mosqueda-Gonzalez BA, Montesinos-López JC, Crossa J, Ramirez NL, et al. A zero altered Poisson random forest model for genomic-enabled prediction. G3 Genes|Genomes|Genetics. 2020;11(2).

[16] Pazhanikumar K, KuzhalVoiMozhi SN. Remote sensing image classification using modified random forest with empirical loss function through crowd-sourced data. Multimedia Tools and Applications. 2024;83(18):53899-921.

[17] Zanatta G, Barroso-Neto IL, Bambini-Junior V, Dutra MF, Bezerra EM, Costa RFd, et al. Quantum Biochemistry Description of the Human Dopamine D3 Receptor in Complex with the Selective Antagonist Eticlopride. Journal of Proteomics & Bioinformatics. 2012;5:155-62.

[18] Wright SN, Kochunov P, Mut F, Bergamino M, Brown KM, Mazziotta JC, et al. Digital reconstruction and morphometric analysis of human brain arterial vasculature from magnetic resonance angiography. NeuroImage. 2013;82:170-81.

[19] Sumwiza K, Twizere C, Rushingabigwi G, Bakunzibake P, Bamurigire P. Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked. 2023;41:101316.

[20] Senan EM, Abunadi I, Jadhav ME, Fati SM. Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms. Computational and mathematical methods in medicine. 2021;2021:8500314.

[21] Kuriakose D, Xiao Z. Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. International Journal of Molecular Sciences. 2020;21:7609.