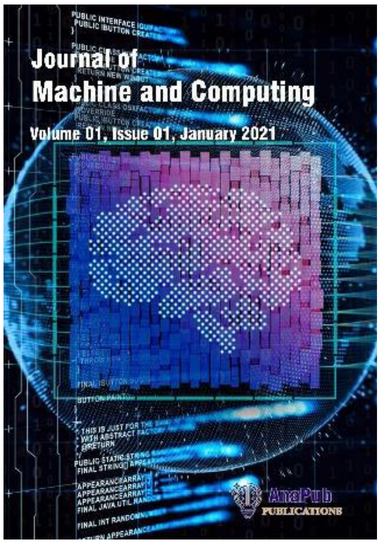
Journal Pre-proof

An Ensemble Cognitive Model for Stroke Prediction Using Unstructured Health Information Powered by Machine Learning



Hayder M. A. Ghanimi, Akilandeswari K, Hanumat Prasad A, Sudhakar Sengan, Badde Praveen Prakash and Ravi Kumar Bommisetti

DOI: 10.53759/7669/jmc202505048

Reference: JMC202505048

Journal: Journal of Machine and Computing.

Received 12 June 2024

Revised form 21 October 2024

Accepted 18 December 2024

AnaPub
PUBLICATIONS

An Ensemble Cognitive Model for Stroke Prediction Using Unstructured Health Information Powered by Machine Learning

Hayder M. A. Ghanimi[1,2], K. Akilandeswari[3,*], A Hanumat Prasad[4], Sudhakar Sengan[5], Badde Praveen Prakash[6], Ravi Kumar Bommisetti[7]

[1]Department of Information Technology, College of Science, University of Warith Al-Anbiyaa, Karbala, Iraq.

[2]Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq. *Email : hayder.alghanami@uowa.edu.iq*

[3]Department of Computer Science, Government Arts College (Autonomous), Salem, Tamil Nadu, 6[illegible]7, India.

Corresponding Author Email : akila.gacslm7@gmail.com

[4]Department of Computer Science and Engineering-Aritifical Intelligence & Machine Learning, Kallam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, Andhra Pradesh, [illegible]22019, India.

Email : hanuma.alahari@gmail.com

[5]Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli–627152, Tamil Nadu, India. Email: sudhasengan[illegible]gmail.com

[6]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, 522302, India. *E-mail: [illegible]de[illegible]een@[illegible]mail.com*

[7]PG Department of Business Administration and Dean, Research and Development Cell, Maris Stella College, Vijayawada, 520[illegible]08, India.

Email : rav[illegible]494[illegible]650@yahoo.com

Abstract

Machine Learning (ML) algorithms have procured a profound position in healthcare sectors, especially in diagnosis, treatments, and recommendation systems. The ML is evolving as an aiding tool for medical practitioners in disease diagnosis. Also, the feature selection reveals the latent relationships among the features, which emerge significant scope for clinical research. In the proposed study, a cognitive ensemble model (CEM) was developed to predict the probability of stroke among various subjects using highly raw clinical data. The optimal base learners are made in such a way that each of them complements one another. The proposed CEM is tested on a real-world dataset on important classification metrics. The results indicate that the CEM deployed in the healthcare sector forewarns patients regarding the probability of stroke.

Keywords: Stroke Prediction, K-Nearest Neighbors, SVM, Random Forest, Ensemble Algorithm, Gaussian Naïve Bayes, CART.

1. Introduction

Stroke occurs when an obstructed condition occurs in the blood vessels, either hindering or bringing down the blood flow to the brain, which is considered a complex human organ. The brain constitutes billions of nerves and blood vessels that collect energy from different body parts. The interrogation of strokes related to the brain is reasonable since it mainly affects the blood vessels that pass to the brain. Hence, stroke is considered to be a cerebrovascular disease. Because the vessels affected by the stroke are accountable for blood flow to the brain, this unexpected cerebrovascular disease is regarded as the second primary source of high mortality and the third largest cause of inability [1].

There are different preventive methods to be followed that can prevent a stroke from happening in the body. Stroke is mainly due to the cause of lifestyle diseases in the body such as uncontrolled blood pressure, carelessness in managing diabetes, and lack of treatment for Heart Disease (HD). Fortunately, these types of health diseases can be addressed to be in control. However, risk factors like age, gender, and heredity can neither be changed nor prevented [2]. The cause of this dreadful disease could be irreversible for many reasons. Hence, stroke prediction at the early stage by assessing the responsible risk factors should be ascertained and alarmed.

Analyzing and predicting stroke risk from the risk factors or symptoms is insufficient to prevent it. Also, perfect decision-making can rarely be achieved by experienced medical experts. With the advanced development in technologies, the opinion by the automated expert system plays a vital role in the prediction of a stroke by closely assessing the sugar level, High and low blood pressure, and other unchangeable risk factors like age, gender, and heredity that give an unprecedented decision to the medical experts. An evaluation of the patients is made in the hospital to take live stroke prediction of the disease at several stages, especially directly from the stroke-affected unit, instead of calculating the patients in the standard unit [3]. Prioritizing care at the early stage of lifestyle disease and other risk factors may reduce the mortality rate to higher levels [4]. The advent of ML has significantly impacted disease prediction in the clinical domain. Many ML models were developed to diagnose the disease, analyze the effect, and propose recommendations, thus reducing the load of medical practitioners.

This work focuses on building an ensemble model for the chances of stroke prediction with four different base learners, namely Support Vector Machines (SVM), K-Nearest Neighbour (K-NN), Gaussian Naïve Bayes (GNB), and Classification and Regression Trees (CART). Random Forest (RF) is used to aggregate the stroke prediction results. The results show that the proposed ensemble algorithm is more accurate in predicting stroke onset. The work also highlights a statistical analysis of the factors contributing to stroke prediction, deriving more insights from the data.

The paper's organization is as follows: Section 2 presents the *state-of-the-art* techniques in stroke prediction. Section 3 shows the knowledge mining activity contributing to stroke and briefs the methods used to enhance the quality of the unstructured clinical data. The proposed CEM is explained in Section 4. The performance analysis of the proposed CEM is validated through important classification metrics, namely accuracy, F1-score, sensitivity, specificity, precision, and recall, and the competitive study is presented in Section 5. Section 6 concludes the work and gives the scope of future research.

2. Related Works

In recent years, the momentum of predictive models in the clinical domain has surged. ML and Deep Learning (DL) models are commonly used in disease diagnosis and treatments by learning the patterns and trends from the clinical data. The predictive models following the classical way are insufficient to handle the dataset in the medical domain [3-6].

The complexity of the problem increases with the dataset; hence, the advent of technologies like DL can be predicted. The attributes that make the stroke prediction are considered the risk factors that share the symptoms' common properties, such as Atrial fibrillation, also called AFib. The current surge in the deployment of many technologies has proved that ML and DL [5] accurately make stroke predictions.

In addition, the combination of ML and pattern recognition [6] is considered one of the denominated methods in stroke prediction complex problems based on neurological diseases, which are regarded as one of the main risk factors for stroke.

Data mining is indulged in mining the patients' symptoms in the available case sheets, taken as the datasets. Standard features are extracted by stemmer [7] from the resultant output and are trained by ML. Out of ML, integrating ML with gradient boosting algorithms gives higher performance and accuracy.

Finding the fundamental reason for the occurrence of the stroke is itself a challenging task. Based on this, it is disclosed that the brain is the primary organ that consumes a large amount of energy from different sources in the body, and the heart inputs the direct power. Unfortunately, any abnormalities in the heart discovered by electrocardiogram cause dysfunctions in the brain, leading to stroke [8].

A significant comparative study was conducted on RF and SVM classifiers, in which the former resulted in higher performance than the SVM [9]. Another hybrid approach in ML highlights the diagnosis of cerebral stroke prediction depending on the physiological data. Even the experts struggle to predict the disease and decide whether to give the treatment based on only the signs unless they are abnormal [10]. Experimental results showed that ML had shown higher performance in predicting the individual functionality of the organs in a human after a stroke [11].

Sometimes, ML may fail to make accurate decisions, which can be solved by ensembling the classifiers to optimize output. Increasing the accuracy of the classification algorithm is as important as predicting the stroke at the early stage, which will decrease the probability of the disease occurring in humans. An ensemble algorithm can handle this to ensure early diagnosis [12].

Recurrent Neural Network (RNN) integrated with a hidden layer is used to analyze the multi-class stroke [13]. Datasets used for stroke analysis are taken from case sheets of patients containing a large amount of clinical data. Label Encoder techniques fill the data, which are lagging in the dataset.

Filling imbalanced information increases the accuracy of classification [14]. Evaluation of Body Mass Index (BMI) relating to mortality rate has a more feasible relationship with the occurrence of stroke prediction [15]. A case study is proposed in the prediction context, and a comparison between novel spiking RNNs and other traditional methods is evaluated [16].

The training dataset for classification by implementing the players not used for testing is validated [17]. Worldwide research was held on stroke prediction at early stages, which could help reduce human disability or mortality due to the disease. An anomaly detection technique is used to detect and assess the health state from the input given by various signals [18]. The detailed comparative analysis is shown in Table 1.

The survey has many implications for creating the dataset using the pre-processing method for imbalanced data. These clinical data are formed based on patients' case sheets for accurate prediction to ensemble the output from the classifiers for higher performance of early diagnosis. Thus, the related works that acquaint the classification of ensemble techniques give versatile diagnosis and prediction at an early stage.

Table 1: Comparative analysis of *state-of-the-art* techniques in stroke prediction

Methodology	Contributions	Limitations
Deep Multi-layered Feed Forward Neural Network	• Comparative analysis of SVM and Naïve Bayes • Stroke prediction percentage	• More risk factors can be involved
Deep Neural Network in Heart Disease Prediction	• A more comprehensive clinical dataset is used. • Comparative Analysis between Logistic Regression and Gradient Boosting Decision tree	• A detailed investigation has to be made in deploying DNN in advanced treatments.
SVM to Predict Stroke	• Structural and functional MRI of the heart are considered as input clinical data. • Lesions and ROI extraction are done.	• More robust algorithms can be designed with collaborative datasets.

Text Mining Tool with ML Classifiers	• Base form generator to obtain input clinical data from web sources • Stemmer to extract root and stem works • Analysis of the impact of stroke on the symptoms of other diseases.	• Only standard classifiers were analyzed.
Dense Convolutional Neural Network	• Stroke prediction from EEG signals • No feature engineering • High accuracy • Reduce subjective perturbation	• Hyperparameter tuning can be done using optimization techniques
Explainable AI	• Preference-based framework • A more informative and explainable learning model • Rule-based metrics	• A completely automated tool can be developed.
DNN with Optimization	• Feature engineering by RF regression • Automated hyperparameter optimization using Auto HPO.	• More detailed analysis can be computed on feature importance. • Physiological indicators can be found.
Ensemble of KNN, SVM, NB, DT, and RF	• Principle Component Analysis and Linear Discriminant analysis-based feature selection • Ensembling of multiple classifiers	• Increases Computational Complexity
Long Short-Term Memory	• RNN predicts with LSTM units	• More factors can be analyzed
DNN with Antlion optimization	• Extensive feature engineering • Optimization algorithms are used to tune the hyperparameters.	• The algorithm can be authenticated on a more robust dataset.
Spiking Neural Network	• Learns spatiotemporal patterns • Reservoir learning • Considers environmental factors	• Hyperparameter optimizations can be done

3. Knowledge Mining from the Dataset

The primary clinical data sources are electronic health records (EHR) and physically measured data from medical centers and wearable devices. Nearly 80% of clinical data is highly unstructured, and tapping those data can reveal new trends and exciting patterns [19]. Improving this unstructured clinical data quality is vital in observing the correlation between biological factors. As stroke is one of the major lifestyle diseases, analyzing its risk factors and controlling them in the early stages will prevent the number of people affected by stroke. The statistical analysis of the dependency between the risk factors concerning age will disclose essential findings. The stroke prediction dataset ideally consists of all the relevant causes of stroke observed among 5110 subjects. Apart from recording the biological parameters, the data consists of information about habits such as smoking status, nature of the job, residence type, and marital status. Table 2 summarizes the fields in the clinical dataset.

Table 2: Description of features in the clinical dataset

Fields	Threshold Level	Mean Value
Patient ID	Unique ID	NA
Gender	{Female, Male}	NA

Presence of Hypertension	{0: No, 1: Yes}	NA
Presence of HD	{0: No, 1: Yes}	NA
Average Glucose Level	[55.12, 271.74]	106.147677
Marital Status	{0: No, 1: Yes}	NA
Work Nature	{Private, Self-employed}	NA
Residence Type	{Rural, Urban}	NA
Smoking Status	{Formerly Smoked, Never Smoked, Unknown, Smokes}	NA
BMI	[10.3, 97.6]	28.893237
Chances of Stroke	{0: No, 1: Yes}	NA

3.1 Pre-processing the data

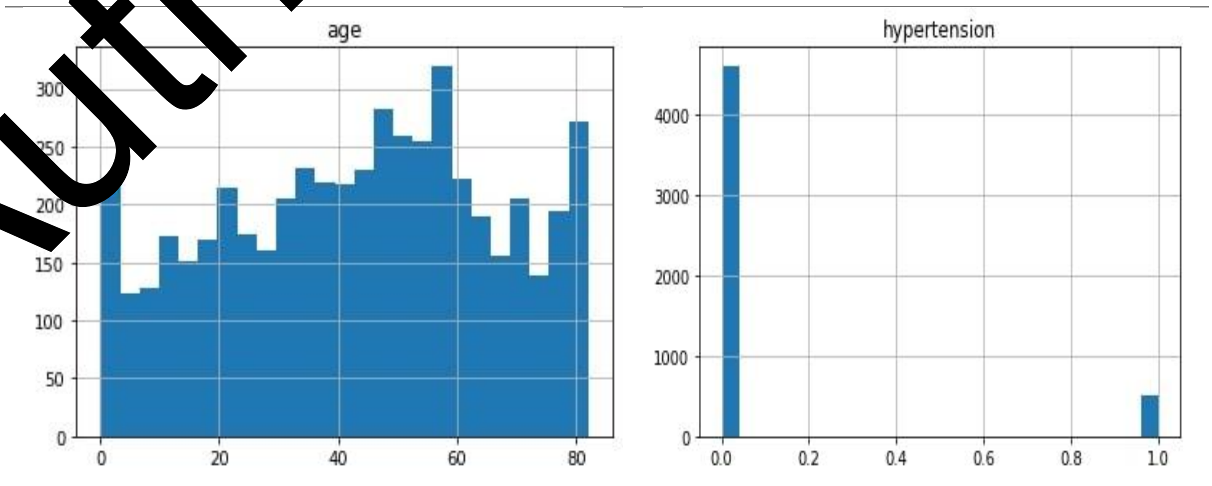
The clinical dataset considered in this study is highly unstructured and needs to be pre-processed. Data cleaning, encoding the definite text data into numerical labels, and filling in missing values are done to improve the data quality.

A. One Hot Encoding: The definite value in the dataset is assigned unique codes for easier processing by ML. In the stroke prediction dataset, the field's gender, marital status, nature of the job, residence type, and smoking status are hardcoded with numerical labels.

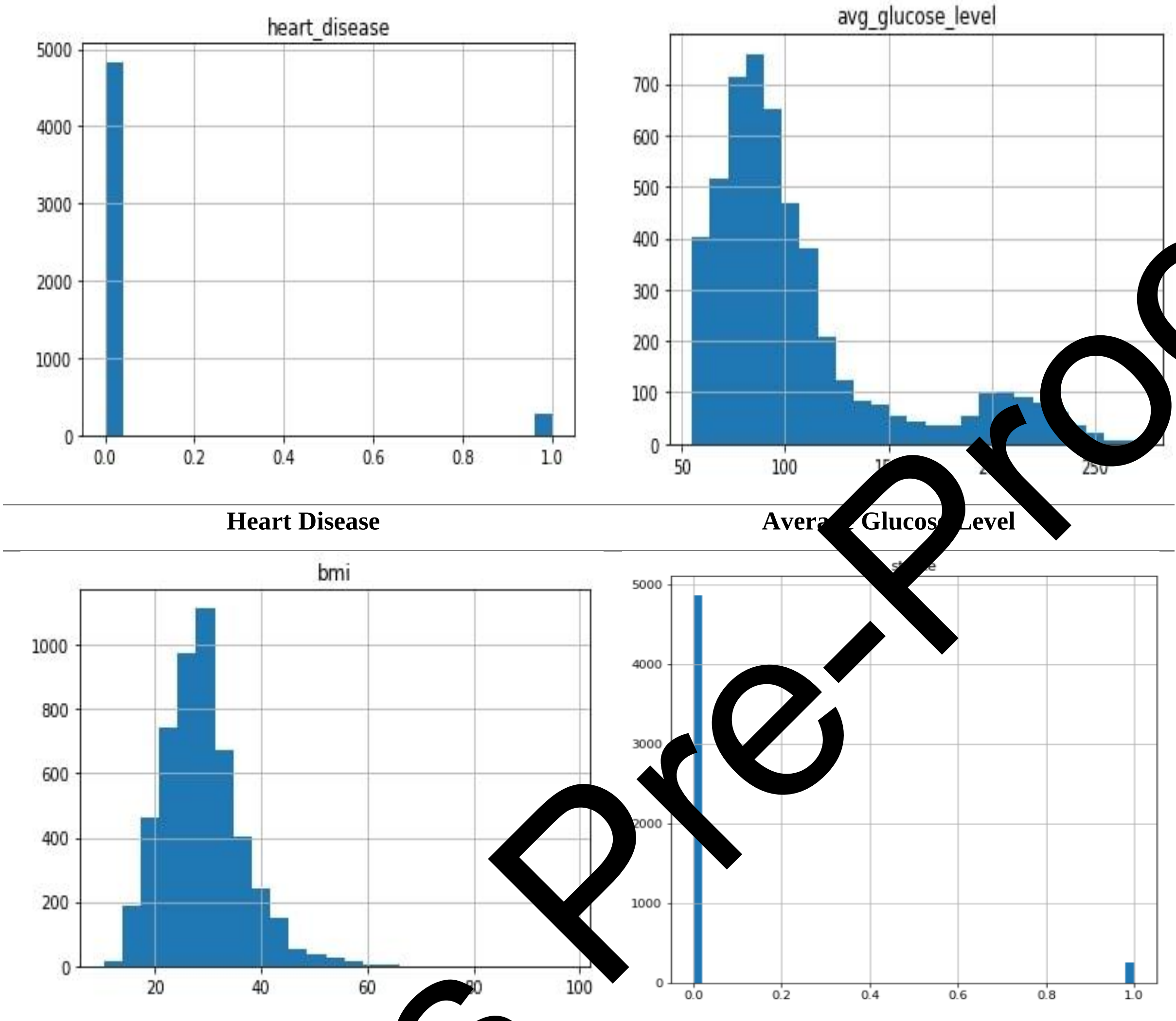
B. Filling Missing Values: Missing values in the fields significantly reduce the predictive power of any algorithm. The dataset under study consists of a handful of missing values in the BMI field. As this is analytical data, the missing values are substituted for their mean value. In case of missing values in the definite fields, they are packed with the mode of the respective fields.

3.2 Exploratory Data Analysis (EDA)

EDA reveals the inherent details of data. This helps build formal models, frame hypotheses, validate the assumptions, and form base work for statistical inferences. The detailed EDA of the stroke dataset is given in Figure 1.



Age	Hypertension



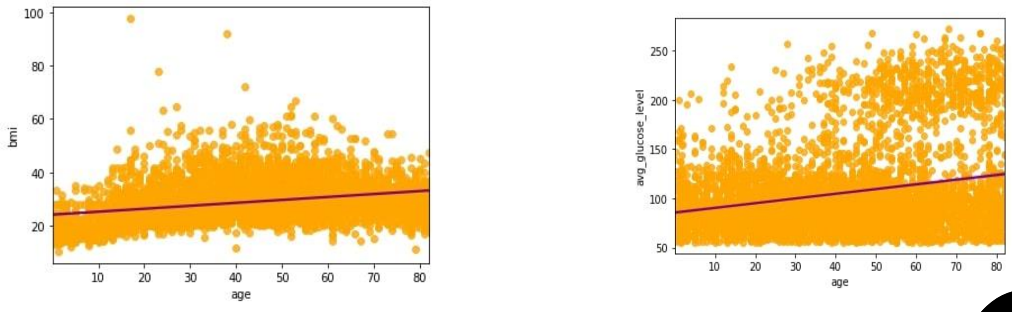
BMI	Stroke

Figure 1: EDA of Factors of Stroke

Figure 1 shows the data distribution and the data composition in the clinical dataset. This analysis indicates that data distribution does not form any pattern, and hence, it is robust. The data collection is done among age groups from 0 to 80+ years. Another notable finding is that the average BMI is between 20-35. The study was conducted on patients with varying glucose levels, as shown in Figure 1. These are some significant inferences that could be drawn from the EDA. A further detailed statistical analysis of these factors concerning stroke will reveal a more lucrative hypothesis, transforming into a significant domain for clinical research.

[illegible]2 Regression Analysis on Factors of Stroke

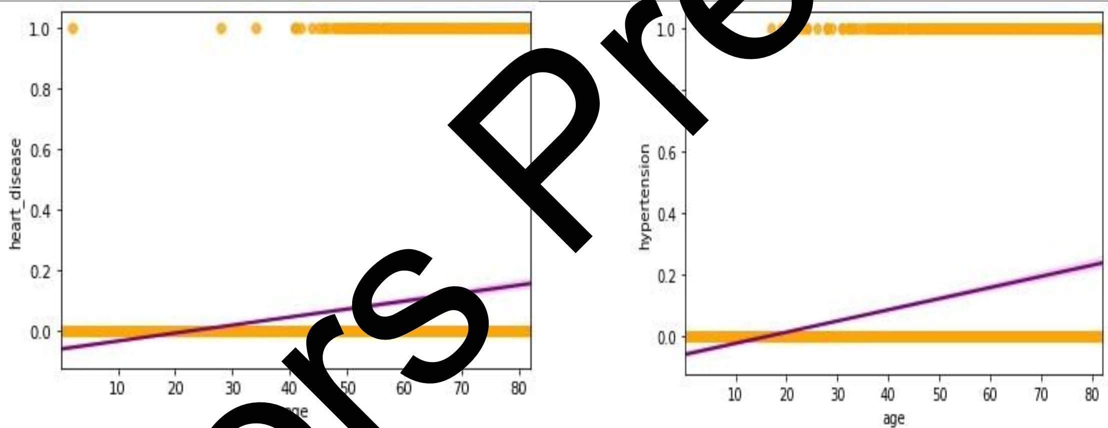
The chances of people developing stroke increases with age. Comorbidities like diabetes, cardiac diseases, and hypertension contribute positively to the long-term disability caused by stroke. The regression analysis of the factors of stroke concerning age will garner attractive benefits such as stroke prediction, delineating the causal relationship between the elements, and predicting the trends between the variables under study.



(a) (b)

Figure 2: Regression Analysis of (a) BMI and Age (b) Glucose Level and Age

Figure 2 (a) and Figure 2 (b) confirm the certainty that the chances of stroke escalate analogously with increased BMI and glucose levels in the older population. The obesity paradox plays a prominent role in increasing the risks of stroke. Also, the primary cause of stroke is damage to the blood vessels, which is contributed by an excess of glucose in the blood.



(a) (b)

Figure 3: Regression Analysis of (a) HD and Age (b) Hypertension and Age

The prolonged presence of cardiac diseases can increase the probability of stroke since the plaque accumulation in the arteries can diminish the oxygen supply to the brain, thus causing the stroke. This precision is confirmed in Figure 3 (a), which shows that the excess strain on the blood vessels due to hypertension weakens the arteries, thus causing the stroke. Figure 3 (b) shows the accelerated risk of stroke in older patients with hypertension. Though the charts display a substantial number of aged patients with hypertension and HD who possess low chances of stroke, the correlation between them demands a more profound investigation.

Detailed analysis of the various factors and their quantitative correlation values are enumerated in Table 3. The other elements, like marital status, gender, and smoking status, did not significantly correlate with the stroke.

TABLE 3: Correlation analysis of stroke with various factors

Factor	Correlation Value
Age	0.2452
HD	0.135
Average Glucose Level	0.132
Hypertension	0.128
BMI	0.0357
Job Nature	0.0064

The summary of stroke prediction among various age groups is shown in Figure 4. This analysis shows that young people in the age group of 35-4[illegible] are also susceptible to stroke—the probabilities of stroke further increase in the elderly population (Figure 4).

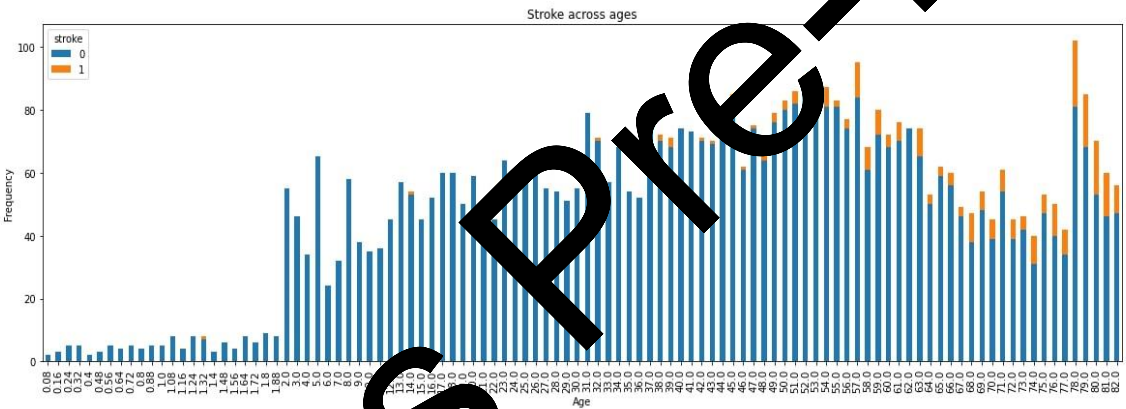


Figure 4: Age-wise Analysis of Stroke

4. CME to Stroke Prediction

As stroke has now slowly evolved as a lifestyle disease, predicting stroke well before its occurrence will be helpful for medical practitioners to forewarn the patients at the onset of early signs. The statistical analysis done in Section 3 elucidates the importance of various factors contributing to the stroke, which forms the features of ML. The diagnosis of diseases such as stroke and cardiac ailments from biological features is crucial. All diagnoses done in the medical field demand high reliability. To ensure this, using decentralized CEM for stroke prediction is always better. The consensus-based approach adopted in building ensemble models substantiates the reliability by accurate prediction. These models are created by combining the power of many homogeneous or heterogeneous-based learning algorithms.

The proposed CEM combines the prowess of k-NN, SVM, CART, and GNB. A more powerful RF classifier smoothens their prediction.

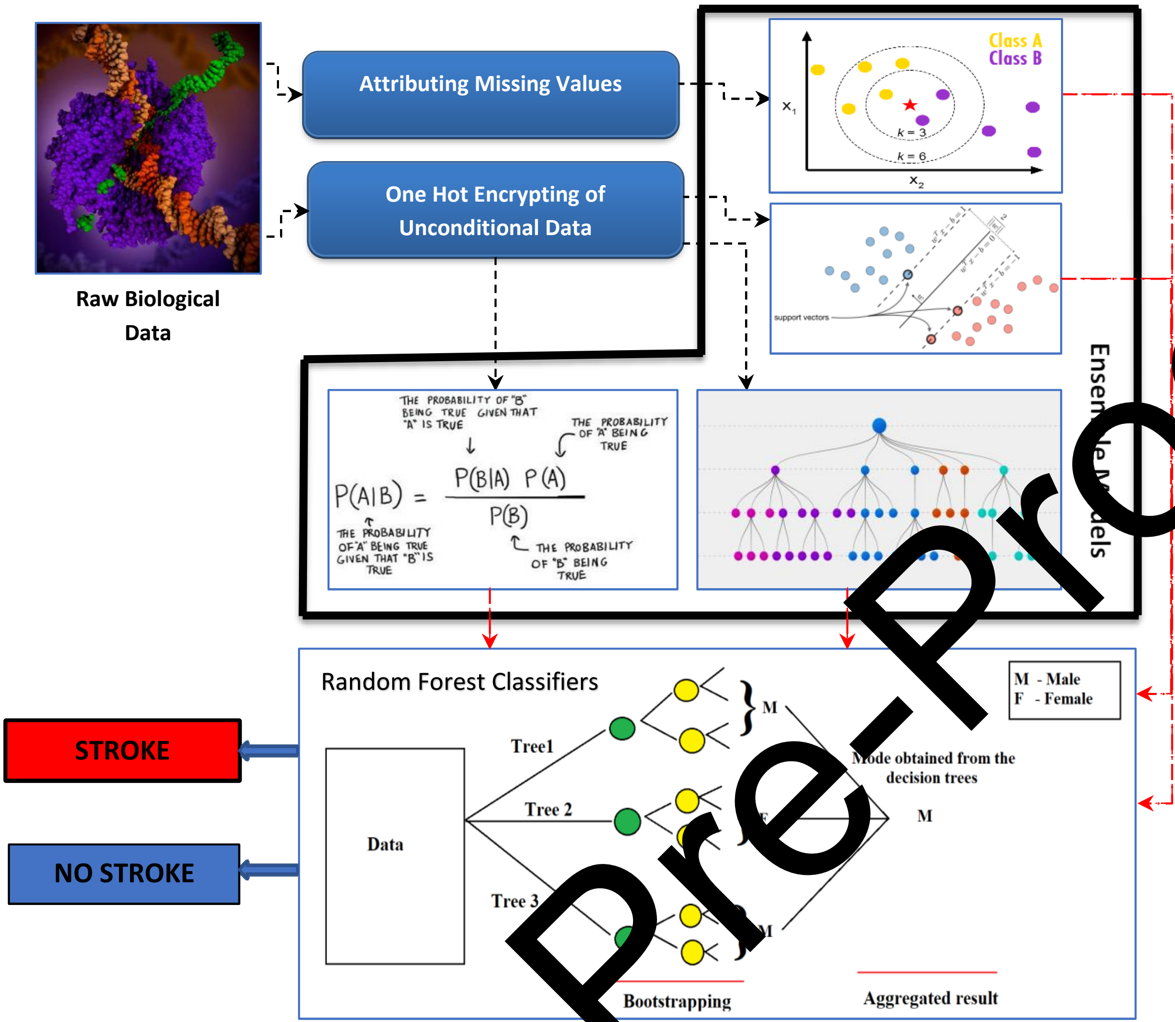


Figure 5: CEM with RF as Meta Classifier

A. Base Learner 1-K-NN: The unsupervised technique classifies the data based on some similarity measures among the data. This non-parametric, lazy learner can effectively classify the noisy data into a predefined number of classes represented as K-value. EQU (1) explains the labelling of data (x) from the pool of data A into an available class (y) based on its probabilistic measure.

$$P(y = j \mid X = x) = \frac{1}{K}\sum_{n \in A} I(y^{(i)} = j) \qquad (1)$$

B. Base Learner 2-SVM: The model ensembles SVM, a small sample learning algorithm that uses structural risk minimization to classify the data. The SVM classifier accomplishes stroke prediction by preserving the linearly separable property of the data. At the same time, the SVM kernels are used to handle the nonlinear data points. The predictive power of the SVM is expressed in EQU (2).

$$y[w^{T}\Phi(x) + \mathrm{bias}] = \begin{cases} \geq 0, \text{ if yes} \\ < 0, \text{ if no} \end{cases} \qquad (2)$$

The term $w^T\Phi(x)+\text{bias}$ refers to the imaginary hyperplane drawn to separate the classes. Thus, SVM is an excellent choice to perform binary classification of data.

C. Base Learner 3-GNB: The classification on Bayes is done independently on the dataset. The GNB classifier is a Bayes algorithm that operates on data typically distributed. This classifier is best for multi-class problems that run on less data. The probabilistic measure of the data belonging to a particular class through GNB is estimated according to EQU (3).

$$P(x,\mu,\sigma)=\frac{1}{\sqrt{2x\pi}}e^{-(x-\mu^2)/2\sigma^2} \tag{3}$$

The estimation of probability is done based on the mean and variance of the normally distributed data points.

D. Base Learner 4-CART Classifier: The CART classifier recursively splits the input data based on the attributes until a proper class is formed. These trees take the dependent variables with a finite number of unordered or continuous data. The performance of the trees is measured in terms of misclassification costs. The proposed model uses the Gini index to partition the data values given in EQU (4).

$$\text{Gini }(X)=1-\sum_{i=1}^{n}p_i^2 \tag{4}$$

P_i is the probability of the set of data X that belongs to a particular class.

E. Metaclassifier-Random Forest: The meta classifier in the CEM predicts the outcome by considering the predictions of the individual base learners as meta-features. The proposed CEM used the RF as a meta classifier as it is another decision tree ensemble. The class label y_i is determined from EQU (5).

$$y_i=\begin{cases}1 \text{ if } p_i>0.5\\ 0 \text{ otherwise}\end{cases} \tag{5}$$

The average probability p_i of individual trees T is computed from the majority voting mentioned in EQU (6).

$$p_i=\frac{1}{T}\sum_{t=1}^{T}I(y_{it}=1) \tag{6}$$

RF can quickly spawn among individual trees, so it is suitable for handling more essential data and deploying RF. This is because the meta-classifier induces randomness in selecting the meta-features from the base learners, thus mitigating the impact of overfitting. The genericity of the RF to be extended to multi-class problems attracts many models to be built using RF.

5. Experimental Analysis of CEM in Stroke Prediction

The dataset's stroke prediction experiment was conducted using the test-train ratio of 70-30. The model is trained on 3397 data with cross-validation K as 10. The following are the performance metrics based on which the assessment of the proposed CEM is presented:

A. Classification Accuracy: It is the rate of correctly classified data. It is the ratio between the number of correctly classified data and the total classifications made. The mathematical formulation of classification accuracy is given in EQU (7).

$$\text{Accuracy} = \frac{\text{Number of instances of rightly classified as stroke and non stroke}}{\text{Total classifications}} \quad (7)$$

B. Specificity: This is the statistical outcome of the True Negatives (TN); that is, the patients predicted to be unaffected by stroke are not prone to stroke. The expression for specificity is given as EQU (8).

$$\text{Specificity} = \frac{\text{Number of instances rightly classified to be not prone to stroke}}{\text{Actual number of healthy patients}} \quad (8)$$

C. Sensitivity: This measure is the statistical outcome of True Positives (TP). Sensitivity is the ratio of people who are predicted to have a probability of being affected by stroke where they are prone to stroke. This test checks whether the model correctly identifies the patients prone to stroke. EQU (9) articulates the expression for sensitivity.

$$\text{Sensitivity} = \frac{\text{Number of instances rightly predicted to be prone to stroke}}{\text{Actual number of stroke affacted patients}} \quad (9)$$

D. Precision: It measures the precision of the model's predictions. The expression for precision is shown as EQU (10).

$$\text{Precision} = \frac{\text{Number of instances that are correctly predicted to be prone to stroke}}{\text{Total number of instances predited to be positive}} \quad (10)$$

E. Recall: Recall is the measure of completeness. In a highly random dataset, recall shares an inverse relationship with precision. When recall increases, the precision may or may not increase, depending on the degree of randomness in the dataset. The mathematical formula for computing recall is given in EQU (11).

$$\text{Precision} = \frac{\text{Number of instances that are correctly predicted to be prone to stroke}}{\text{Total number of instances predited to be positive}} \quad (11)$$

F. F1-score: This is a measure to balance the trade-off between precision and recall, and it is the geometric mean to precision and recall. The expression for the F1-score is mentioned in EQU (12).

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

5.1 Results and Discussions

The proposed CEM is validated by comparing the metrics discussed in Section 5 with base learners and ensembling the base learners with different meta-classifiers. Table 4 enumerates the summary of the results.

Table 4: Performance comparison of individual base learners and various ensemble algorithms

Type of Classifiers	Accuracy	Specificity	Sensitivity	Precision	Recall	F1 Score
GNB	86.2	85.9	85.76	94	86	89
CART	95.1	94.56	94	92	95	93
SVM	94.71	94.3	94.12	[illegible]	95	93
RF	95.5	95	95.5	91	[illegible]	93
Extreme Gradient Boosting (XGB)	95.5	60	95.69	[illegible]	96	94
KNN + SVM + GNB + CART Metaclassifier: Logistic Regression (LR)	96.3	80	81	95	97	96
K-NN + SVM + GNB + CART Metaclassifier: GNB	96.79	81	82	95	97	96
K-NN + SVM + GNB + CART Metaclassifier: RF	**[illegible]**	**86**	**85**	**96**	**98**	**97**

The detailed analysis of various ML in stroke prediction shows that the proposed CME with the RF as a classifier shows improved performance over the other models. The graphical analysis of the same is depicted in Figure 6. The efficacy of the proposed CEM on 70% training is a positive note, as the model can predict the stroke rate with substantially less training.

Further, the accuracy can still be raised by scaling up the data and including more attributes for upgraded predictions.

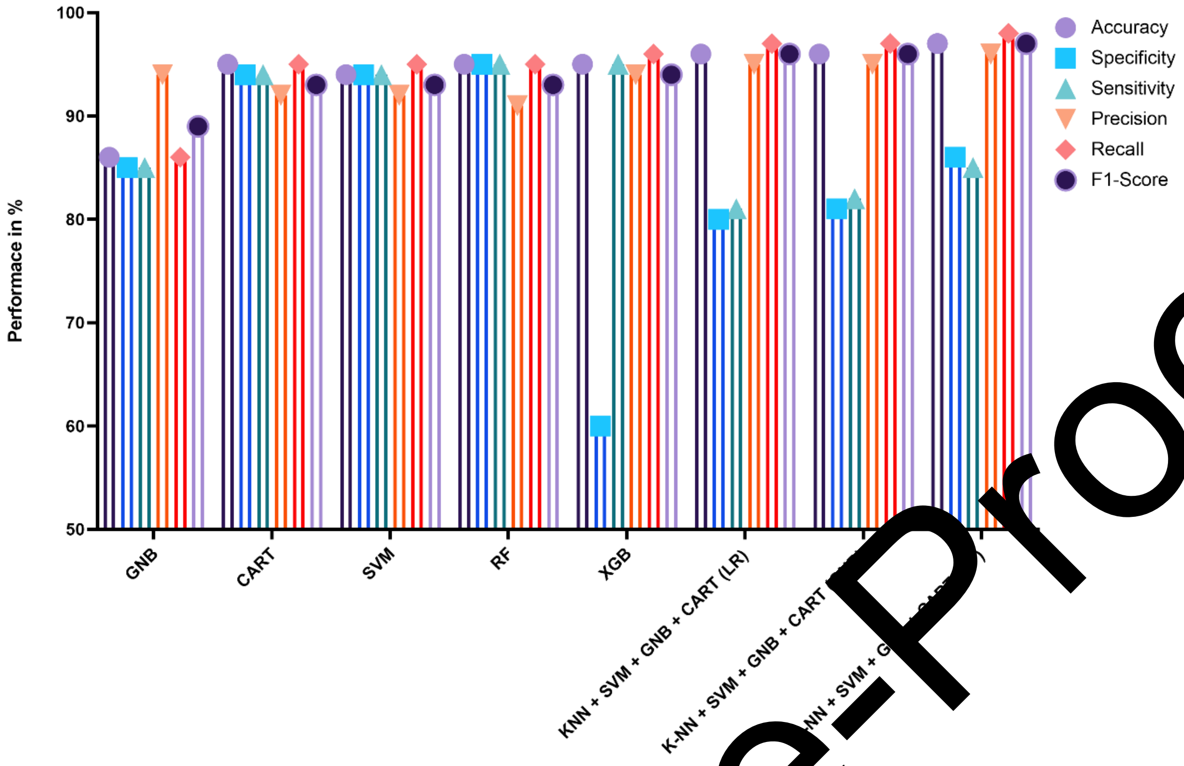


Figure 6: Performance Analysis of ML in stroke prediction

6. Conclusion and Future Work

This article focuses on correlation analysis of various factors of stroke to unveil the relationship among them. The proposed CEM integrates the predictive power of SVM, KNN, CART, and GNB with RF as a classifier. Each base learner used in the model building has unique strengths, and other base learners complement their inherent weaknesses. The proposed CEM exhibited improved classification accuracy, F1-score, sensitivity, precision, specificity, and recall. The predictive power of the stroke prediction model can be extended by including more attributes.

References

[1] X. Chen, "Stroke Disease Prediction Based on Multi-Model Ensemble Learning," 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 2024, pp. 2129-2132, doi: 10.1109/AINIT61980.2024.10581532.

[2] W. Chen, D. Zhai, H. Wu, F. Liu, Y. Fu, Y. Chen, et al., "Non-contact blood pressure detection based on weighted ensemble learning model", *Signal Image and Video Processing*, vol. 18, no. 1, pp. 553-560, 2024.

[3] N. Hussain, A. Qasim, Z. Akhtar, A. Qasim, G. Mehak, L. del Socorro Espindola Ulibarri, et al., "Stock Market Performance Analytics Using XGBoost", *Lecture Notes in Computer Science*, vol. 14391 LNAI, pp. 3-16, 2024.

[4] F. Tambon, A. Nikanjam, L. An, F. Khomh and G. Antoniol, "Silent bugs in deep learning frameworks: an empirical study of Keras and TensorFlow", *Empirical Software Engineering*, vol. 29, no. 1, 2024.

[5] Rozikin, A. Buono, C. Arif, S. Wahjuni and Widodo, "Classification of the Severity of Downy Mildew Disease Using LGBM", *2023 International Conference on Informatics Multimedia Cyber and Information Systems ICIMCIS 2023*, pp. 364-368, 2023.

[6] M. A. Kumar, G. Manivasagam, K. Kathirvel, V. Kavitha and A. Gupta, "Enhancing the Prediction of Diabetics using Bagging Ensambler Classifier," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 489-495, doi: 10.1109/IITCEE57236.2023.10090921.

[7] S. Georganos, T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, S. Kalogirou, et al., "Is More: Optimizing Classification Performance through Feature Selection in a Very-High-Resolution Remote Sensing Object-Based Urban Application", *GISci. Remote Sens.*, vol. 55, pp. 221-242, 2018.

[8] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Patient-Centric HetNets Powered by Machine Learning and Big Data Analytics for 6G Networks," in IEEE Access, vol. 8, pp. 85639-85655, 2020, doi: 10.1109/ACCESS.2020.2992555.

[9] G. Jain, S. Chopra, S. Chopra and A. S. Parihar, "Attention-Net: An Ensemble Sketch Recognition Approach Using Vector Images," in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 1, pp. 136-145, March 2022, doi: 10.1109/TCDS.2020.3023055.

[10] U. Anwar, S. Khan, T. Arslan, T. C. Russ and P. Lomax, "Radio Frequency-Enabled Cerebral Blood Flow Monitoring and Classification Using Data Augmentation and Machine Learning Techniques," in IEEE Sensors Journal, vol. 24, no. 19, pp. 31040-31053, 1 Oct.1, 2024, doi: 10.1109/JSEN.2024.3444192.

[11] R. W. J. Weijs, D. A. Shkredova, A. C. M. Brekelmans, D. H. J. Thijssen and J. A. H. R. Claassen, "Longitudinal changes in cerebral blood flow and their relation with cognitive decline in patients with dementia: Current knowledge and future directions", *Alzheimer's Dementia*, vol. 19, no. 2, pp. 532-548, Feb. 2023.

[12] L. Zeng et al., "A non-invasive and comprehensive method for continuous assessment of cerebral blood flow pulsation based on magnetic induction phase shift", *PeerJ*, vol. 10, Feb. 2022.

[13] S. J. van Bohemen, J. M. Rogers, P. C. Boughton, J. L. Clarke, J. T. Valderrama and A. Z. Kyme, "Continuous non-invasive estimates of cerebral blood flow using electrocardiography signals: A feasibility study", *Biomed. Eng. Lett.*, vol. 13, no. 2, pp. 185-195, May 2023.

[14] U. Anwar, T. Arslan, A. Hussain, T. C. Russ and P. Lomax, "Design and evaluation of wearable multimodal RF sensing system for vascular dementia detection", *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 5, pp. 928-940, Oct. 2023.

[15] U. Anwar, T. Arslan, A. Hussain and P. Lomax, "Wearable RF sensing and imaging system for non-invasive vascular dementia detection", *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 4-8, May 2023.

[16] Z. Gong et al., "Dynamic cerebral blood flow assessment based on electromagnetic coupling sensing and image feature analysis", *Frontiers Bioengineering Biotechnol.*, vol. 12, pp. 1-14, Feb. 2024.

[17] N. N. Nisha et al., "A deep learning framework for the detection of abnormality in cerebral blood flow velocity using transcranial Doppler ultrasound", *Diagnostics*, vol. 13, no. 12, pp. 2000, Jun. 2023.

[18] L. Cai et al., "A machine learning approach to predict cerebral perfusion status based on internal carotid artery blood flow", *Comput. Biol. Med.*, vol. 164, Sep. 2023.

[19] D. J. Vitello, R. M. Ripper, M. R. Fettiplace, G. L. Weinberg and J. M. Vitello, "Blood density is nearly equal to water density: A validation study of the gravimetric method of measuring intraoperative blood loss", *J. Veterinary Med.*, vol. 2015, pp. 1-4, Jan. 2015.