WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# A Survey on Ensemble Learning Approach for Apoplexy Prediction

***Mitali Kadam[1], Poonav Kuchekar[2], Srushti Deopurkar[3], Sunita Naik[4]***

*[1, 2, 3]Final Year Student, Computer Engineering Department, VIVA Institute of Technology, India*
*[4]Assistant Professor, Computer Engineering Department, VIVA Institute of Technology, India*

**A B S T R A C T**

In recent years stroke are one of the leading causes of death by affecting the central nervous system. The term apoplexy refers to brain stroke. There are different types of strokes, among which ischemic and hemorrhagic majorly damages the central nervous system. Brain stroke is a critical medical condition that requires immediate attention to prevent further damage to the brain. Early detection of the risk factors associated with brain stroke can help in timely intervention and prevention of this condition. In this context, the use of machine learning algorithms for brain stroke prediction has gained significant attention in recent years. Multilevel stacking is a powerful technique that combines the outputs of several machine learning algorithms to improve the overall performance of the prediction model. In this research work, Machine learning techniques are applied in identifying, classifying and predicting the brain stroke from medical information. The standard dataset is available on Kaggle. The dataset contains 11 attributes and 5000 rows. The stacking was composed of single classifier as base learners and Logistic Regression or Random Forest was used as meta learner. Decision Tree, K-Nearest Neighbor, SVM, Random Forest, Logistic Regression etc was used as a single model or as an ensemble model.

Keywords: Base learners, Brain Stroke, Decision Trees, Ensemble learning, Logistic Regression, Meta-learner, Stacking.

## 1. Introduction

Brain stroke is a medical condition that has become one of the leading causes of death in the world, according to the World Health Organization (WHO). There are two main types of brain stroke: hemorrhagic and ischemic. Hemorrhagic stroke occurs when blood vessels rupture due to hypertension, while ischemic stroke occurs when the blood flow to the brain decreases or stops altogether. Both types of stroke can cause parts of the brain to stop functioning properly, leading to a range of symptoms such as the inability to feel one side or movement of one side. The main causes of brain stroke include hypertension, heart disease, obesity, and diabetes. Machine learning (ML) is a field of artificial intelligence that utilizes data and algorithms to make predictions based on features from a given dataset. ML continues to evolve and has become increasingly beneficial in programming tasks that can predict or classify data. Classifiers separate data into classes, and prediction functions create a trend line, also known as the line of best fit, to fit a shape that gets closest to the data points. ML can fall into three categories: supervised machine learning, unsupervised machine learning, and semi-supervised learning. The classifiers used in this study fall under supervised and unsupervised machine learning. An ensemble model is a valuable ML algorithm that can provide a variety of techniques for classification and regression. This study focuses on classification techniques.

There are several ensemble techniques available, such as bagging, boosting, stacking, and blending. The stacking ensemble model creates a strong meta-classifier, which is trained on features that are outputs from the combination of weak or base-level classifiers. With the use of ML, algorithms are trained to find patterns using a large dataset. ML can fall into three categories: supervised machine learning, unsupervised machine learning, and semi-supervised learning. The classifiers used in this study fall under supervised and unsupervised machine learning. An ensemble model is a valuable ML algorithm and can provide a variety of techniques for classification and regression. This study focuses on classification techniques. There are several ensemble techniques available such as bagging, boosting, stacking, and blending. The Stacking ensemble model creates a strong meta-classifier, which is trained on features that are outputs from the combination of weak or base level classifiers. With the use of ML, algorithms are trained to find patterns using a large dataset to make predictions. Some past researchers have used ML models to predict brain stroke using single classifiers and have had some high success rates. Stacking combines base learners/weak learners into a more robust model than individual base learners. With a right combination of weak or base learning algorithms, a meta-model with lower bias and variance can be developed. Base learners are trained in parallel and combined by training a final meta-model on the output predictions using different weak models. In research, the base learner used was; Support Vector Classifier, Gradient Boost, Random Forest, Decision Tree, K-Nearest Neighbor. and Logistic Regression is used as Meta learner or meta model. Also, Logistic Regression proved to be the best meta-learner on every dataset mostly as it maps the input correctly generated by one or more model to give final prediction.

## 2. Related Work

Chirag Rana, et.al[1] have used SMOTE (Synthetic Minority Oversampling Technique) to predict the correct values of having a brain stroke. In dataset there are most values of true positives rather than false positives, false negative and true negative. It can only correctly study or analyze the value of non-

strokes patients. SMOTE is a technique which is applied on training dataset rather than testing dataset. Testing dataset needs to remain untouched. As the values of dataset are categorically so the author applied Label encoder on 'gender, ever_married and Residence_type' and One Hot Encoder on 'Work_type and smoking_status'. The author also proposed a Neural Network model with two hidden layer of 50 and 20 neurons. Relu activation is used in hidden layers. The Sigmoid Function and Stochastic Gradient Descent optimizer is also used. The accuracy is 86% with ROC-Curve of 0.843.

Samaa Ahmed Mostafa, et.al[2] research based on survey on ensemble models are conducted in this paper. By researching about 30 papers, the stacking ensemble algorithm proved to be the best. The difference is shown between single classifier and ensemble classifier. SVM has accuracy of 77.29%, KNN has accuracy of 84.58%, Decision tree 86.10%, Random Forest 86.63%, AdaBoost 82.43% and Stacking 87.58%. The stacking with Neural network was 95% accuracy.

Minhaz Uddin Emon, et.al[3] A weighted voting classifier is used. Ten classifier were used. Logistic Regression (LR), Stochastic Gradient Descent (SGD), Decision Tree Classifier (DTC), AdaBoost, Gaussian, Quadratic Discriminant Analysis (QDC), Multi-Layer Perceptron (MLP), K-Neighbors, Gradient Boosting Classifier (GBC), XGBoost (XGB). The accuracy of weighted voting was 97%. The ten classifiers input is taken as output by Weighted voting classifier then the model optimization is done based on confusion matrix, AUC, ROC, F1 score, recall, precision FP rate and FN rate.

Kunder Akash Mahesh, et.al[4] three methodologies are used to predict brain stroke on dataset 'Decision tree, Naïve Bayes, Artificial Neural Network'. Decision Tree is one of the algorithm which handles high dimension data. These tree based conditional methods empower predictive models with high accuracy. Naïve bayes is used in sentimental analysis as it is a probabilistic machine learning model. It is mostly used in spam filtering, recommendations etc. Neural network are modeled as human brain. They sense data and can interpret it through clustering raw input or by machine perception. Out of three methodologies, Neural network show acceptable accuracy for stroke patients.

Elias Dritsas, et.al[5] In this research work, the Stacking algorithm is used. Stacking consist of J48, Reptree and Random Forest Algorithms. Logistic Regression was used as meta learner. Through research the author used Naïve Bayes, Logistic Regression, 3-Neural Network, Stochastic Gradient Descent, Decision Tree(J48), Multi-layer perceptron, Majority Voting, Random Forest and Stacking. Stacking has accuracy of 98.9% with the precision, f1-score and recall of 97.4%.

José Alberto Tavares, et.al[6] In this research work, the CRISP-DM methodology is used. Cross Industry Standard Process of Data Mining. The CRISP-DM methodology includes Business Understanding,

Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The dataset has 5110 rows and 11 attributes. SMOTE is used to balance the imbalanced dataset. According to exploratory data analysis, the hypertension and heart disease share common relationship with stroke data. That means when the individual has both of them as positive then there are higher chances of stroke. The different ML models used to analyze the stroke; Decision Tree, Logistic Regression, Naïve bayes, K-Nearest Neighbor, Random Forest, Neural Network, Support Vector Machine, XGBoost. Random Forest performed great with the accuracy of 92.32%. It predicted 70 false negatives and 96 false positives with the 0.917 and F-score 0.92.

Tahia Tazin, et.al[7] Conducted a comparative study on Random Forest, Logistic Regression, Decision Tree and Voting classifier. According to research, Random forest did exceptionally well than voting classifier with an accuracy of 0.96 while the accuracy of voting classifier is 0.91. Logistic Regression has very low accuracy 0.79. Decision Tree also did well than Voting classifier with the accuracy of 0.94.

Jiayuan Ling, et.al[8] The dataset used for We-Chat application was 1915 initiators, 37348 actors and 57053 rows for logging with different IP address. Data was converted into minute integers in order to have consistency. Two Level Stacking Model is used to predict abnormal activities. The logistic regression wasn't stable enough so SVM rbf kernel is used as meta model. The accuracy of two level stacking model is 0.80.

Shailendra Singh, et.al[9] This research introduces a multi-level stacking ensemble learning mechanism for accurate electricity load forecasting in the Internet of Energy ecosystem. The architecture consists of four layers utilizing strong learners such as Random Forest, Cubist, k-Nearest Neighbors, Xtreme Gradient Boosting, Support Vector Machine Regression, Multivariate Adaptive Regression Splines, and Principal Component Regression. The proposed approach includes both original features and meta-features extracted during stacking to improve prediction accuracy.

Abd Mizwar, et.al[10] The purpose of this research is to increase accuracy by proposing the application of one of the ensemble learning algorithms, namely the Extreme Gradient Boosting algorithm. After the previous process is complete, the classification process uses the Extreme Gradient Boosting method and continues for the evaluation process using a confusion matrix to measure model performance in the form of precision accuracy, and recall.

Muath A. Obaidat, et.al[11] This study's use of a diverse ensemble model, combining three base-level classifiers, to predict heart disease risk and lower global mortality rates is a unique and effective approach, with superior performance compared to individual classifiers. Evaluation metrics include accuracy, AUC, specificity, precision, and sensitivity, and the methodology was implemented using open-source software, R-Studio.

Lesi Nnadozie, et.al[12]This research aims to create a multi-level stacking ensemble model to predict real estate prices using data from Port Harcourt real estate firms. The model uses Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, Decision Tree Regression, and ElasticNet Regression algorithms. The best performing stacked models were combined to create the final model, which outperformed individual base models with an R-square of 0.953205, MSE of 0.001, RMSE of 115923, MAE of 0.065477, and training time of 0.39919.

Shivani Batra, et.al[13] This research proposes an ensemble imputation model that is educated to use a combination of simple mean imputation, k-nearest neighbor imputation, and iterative imputation methods, and then leverages them in manner where ideal imputation strategy is opted based on attribute correlations on missing value features. We introduce a unique Ensemble Strategy for Missing Value to analyze healthcare data with considerable missing

values to identify unbiased and accurate prediction statistical modeling. The performance metrics have been generated using Xtreme gradient boosting regressor, random forest regressor, and support vector regressor. The study uses real-world healthcare data to conduct experiments and simulations of data with varying feature-wise missing frequencies indicating that the proposed technique surpasses standard missing value imputation approaches as well as approach of dropping records holding missing values in terms of accuracy.

Pradeep Kumar Kushwaha, et.al[14] This paper compares the proposed GA-SLE system with various machine learning algorithms using 10-fold cross-validation and hyper-parameter tuning. The analysis includes a regression comparative study and shows that GA-SLE outperforms other classifiers with a prediction accuracy of 99.8% and minimum error loss. The system enables early detection and diagnosis of sickness, making it a valuable tool for practitioners.

Nourhan M Swelam, et.al[15] This study compares the performance of individual supervised learning classification algorithms with a stacking ensemble learning approach using breast cancer dataset. The proposed stacking model outperforms individual algorithms with 99.08% accuracy for diagnosis and 98.05% accuracy for prognosis.

## 3. Research Objectives

Random Forest and Stacking provides an efficient solution in predicting the brain stroke. The multilevel stacking refers to where multiple layers of weak learner also called as meta-learner are stacked. According to research Random Forest, Decision Tree classifier, K-Nearest Neighbor, Gradient Boost, Support Vector Machine and Logistic Regression works accurately on the stroke dataset. This research aims to develop a model with all this base model and create a strong model using two-level multilevel stacking.

## 4. Proposed System & Methodology

### *4.1 Algorithm for Multilevel Stacking for predicting the chances of stroke.*

Step 1: Import all necessary packages.

Step 2: Import Brain Stroke dataset from kaggle.

Step 3: Clean dataset, find missing values.

Step 4: Visualize the dataset.

Step 5: Balance the imbalanced dataset using SMOTE.

Step 6: Apply Label Encoder on dataset.

Step 7: Use 80% of data for training and 20% for testing.

Step 8: Calculate the accuracy using Random Forest, Logistic Regression, Decision Tree, K-Nearest Neighbor, Gradient Boost, Support Vector Classifier.

Step 9: Stack the Random Forest and SVC with meta-learner as Decision Tree.

Step 10: Stack the Random Forest and GB with meta-learner as KNN.

Step 11: Take the inputs from Decision Tree and KNN and combine them through meta-learner as Logistic Regression.

Step 12: Calculate the Accuracy.

Step 13: Import the pickle file named ‘model.pkl’.

### *4.2 Algorithm for Predicting Time Constraint for Stroke.*

Step 1: Import all necessary packages.

Step 2: Import the stroke dataset

Step 3: Clean the dataset.

Step 4: Drop some columns like gender, ever married, residence type etc.

Step 5: Balance the dataset using SMOTE.

Step 6: Apply Label Encoding.

Step 7: Train and Test the data

Step 8: Calculate the accuracy using Random Forest.

Step 9: if hypertension ==1 and heart disease ==1 and glucose level > 150 and smoke_smokes==1: print("YOU HAVE HIGH CHANCES OF STROKE, CONSULT DOCTOR ASAP") else if hypertension==1 or heart disease==1and glucose level >130 and smoke_smokes==1: print ("YOU HAVE CHANCES OF STROKE WITHIN COMING YEARS") else: print("YOU HAVE NO CHANCE OF STROKE")

### *4.3 Algorithm for Jupyter Notebook to Flask*

Step 1: Import all necessary packages.

Step 2: import model.pickle file

Step 3: Route app to index.html

Step 4: Create form.

Step 5: Predict data on the basis of given input.

Step 6: Predicted data is shaped into a numpy array and then is calculated.

Step 7: All index files are rendered from templates folder.

### *4.4 Data Collection:*

The production of the project started by collection of dataset. The dataset consists of 5110 rows and 12 attributes. The attributes are gender, age, hypertension, heart disease, ever married, Residence type, average glucose level, body mass index, smoking status, work type etc. The dataset is collected from Kaggle which is a standard dataset. By analyzing data there were 201 missing values of BMI. To find this missing value fillna() method is used in. this project. Exploratory data analysis gives us insights about the data in the dataset. According to exploratory data analysis the dataset is highly imbalanced and thus is required to balance the dataset first. SMOTE is used to balance the dataset. SMOTE is an improved method of dealing with imbalanced data in classification problems.

### *4.5 UI Architecture:*

The Fig (1) depicts the architecture of system, the flow of modules from flask to web application. The app.py is the main file which contains the routes to all html files and it fetches the data from the model.pkl file.

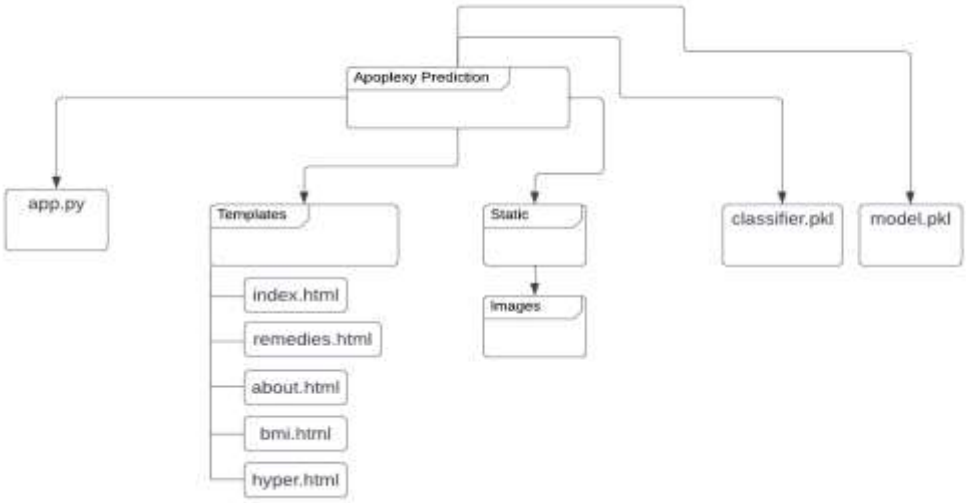
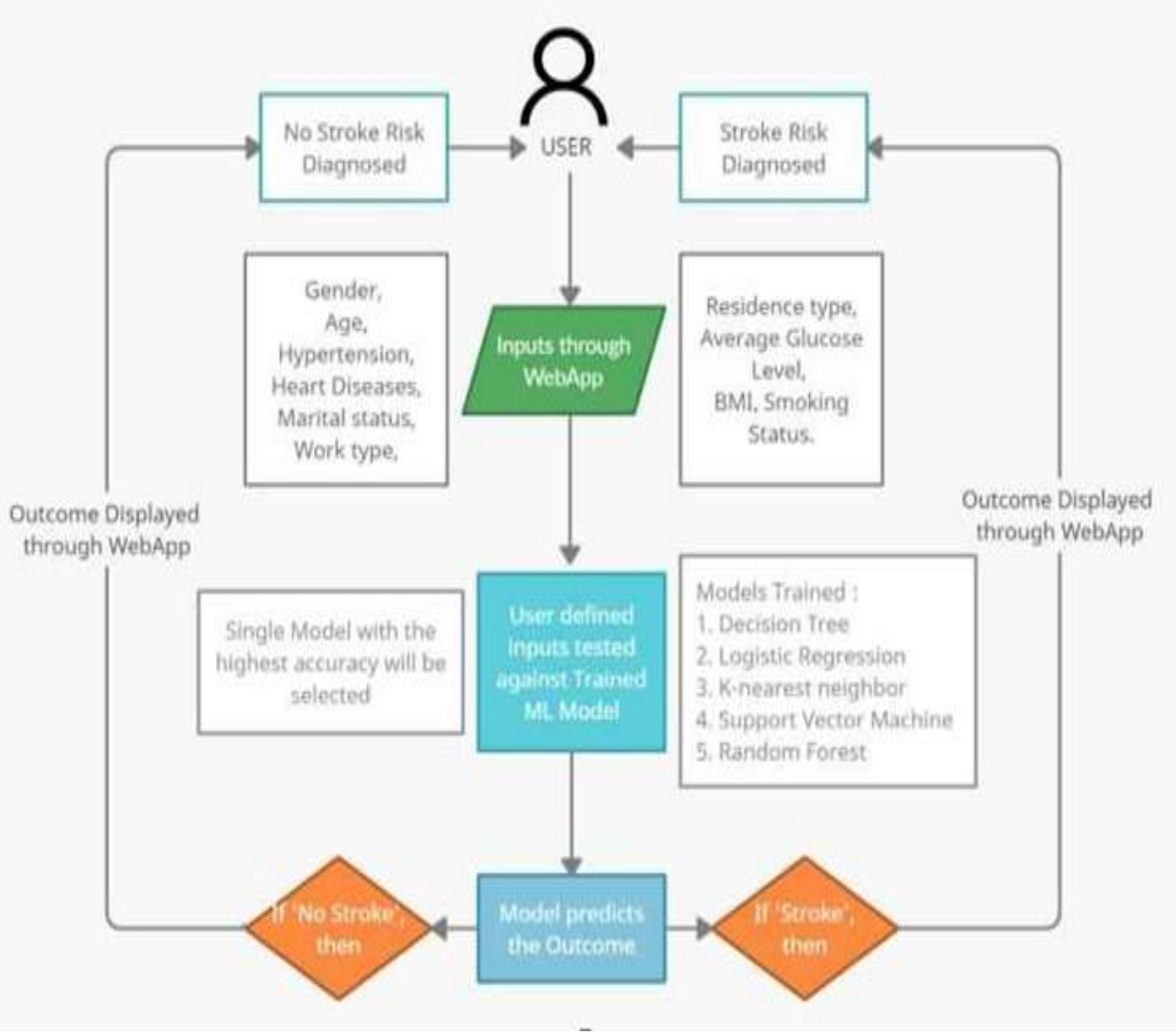



**Fig. 1 UI Architecture**

### *4.6 UI Workflow Architecture*

The Fig (2) Shows that the person using our Web Application will be the user who wants to know whether they have a risk of having Brain or not. The user will be asked about some details regarding their gender, age, hypertension, heart diseases, marital status, work type, residence type, average glucose level, BMI and smoking status. All these details are necessary for the prediction of stroke possibility for that individual. Through our Web Application,

the user will get to know about the outcome of its input data. The outcome for “No Stroke” will be displayed as “No Stroke Risk Diagnosed” otherwise,




the user will get to know about the outcome of its input data. In the case for “Stroke” as an outcome, it will be displayed as “Stroke Risk Diagnosed

**Fig.2 UI Workflow Architecture**

***4.7 Architecture Of System:***

The Fig (3) depicts architecture of system i.e algorithm used in the system. Multilevel stacking refers to stacking in levels of the meta classifier. The meta classifier will take the output of other models as input and then combine it together. Here the meta classifier are Decision Tree, K-Nearest Neighbor and Logistic Regression. The multi-level stacking architecture consists of 5 base classifiers. According to research, Random Forest prove to be the best classifier.

Also, Decision Tree, Gradient Boost, Support Vector Classifier, K-Nearest Neighbors and Logistic Regression works best on this dataset. The k-fold cross validation is used to avoid overfitting. As k-fold is better than handout method to handle overfitting. Logistic Regression maps the inputs of all outputs accordingly so it is used as meta-classifier.

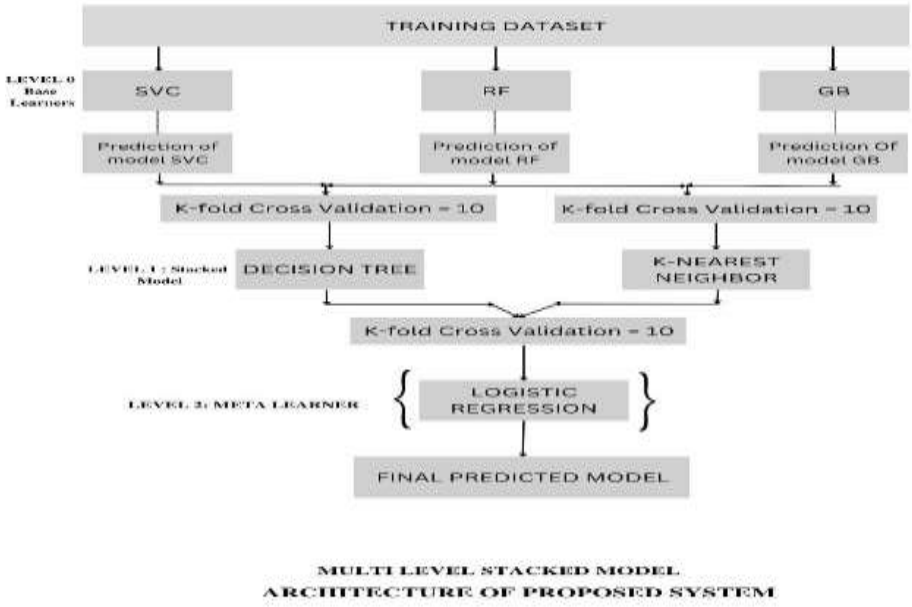



**Fig. 3 Proposed System Architecture**

***4.8 General Block Diagram:***

Firstly, the data is collected and then it is divided into training and testing dataset i.e it is splitted by 70-30% or 80-20% or 50-50%. Then the training data is trained using several algorithms and then evaluation is done with the help of an algorithm on testing dataset and then final model is built.

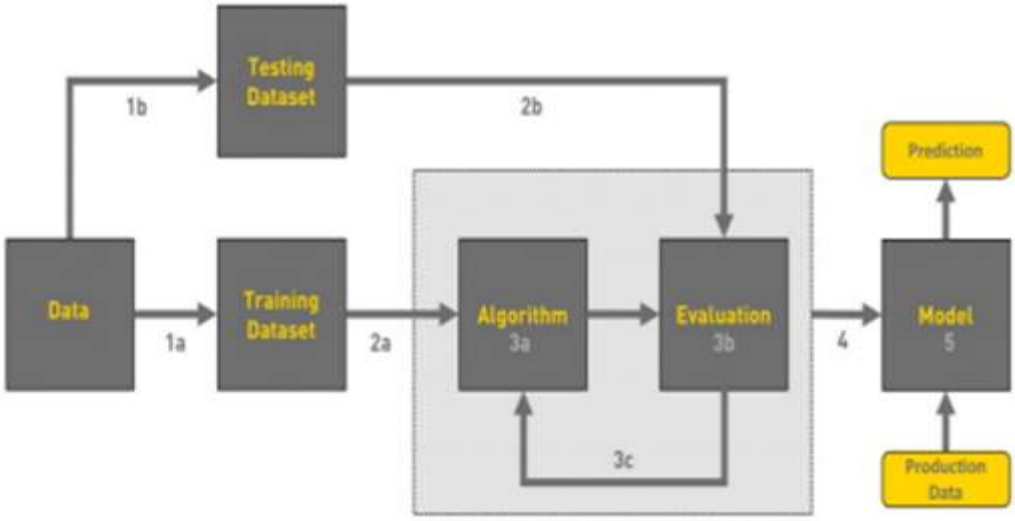


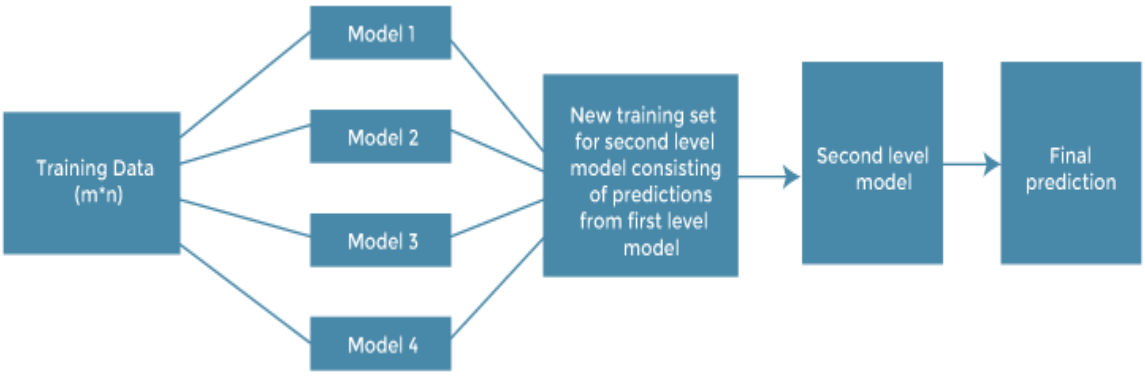

**Fig. 4 Block diagram of Proposed System**

## 5. Methodology:

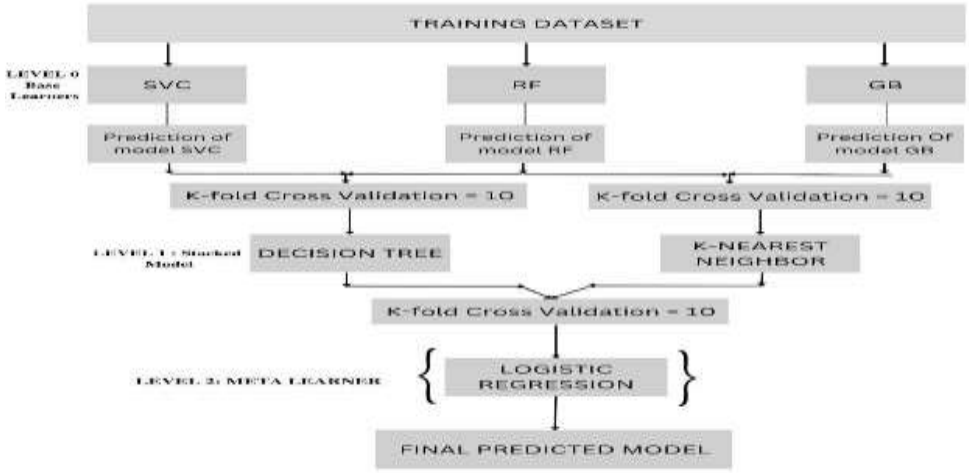
Production of the system begins with processing the dataset. The dataset is from kaggle. The dataset has 5110 rows and 11 attributes. The attributes are age, gender, hypertension, heart disease, bmi, average glucose level, smoking status etc. The dataset has 201 missing values of BMI and KNN imputer technique is used to find the missing values of BMI. KNNImputer by scikit-learn is a widely used method to impute missing values. It is widely being observed as a replacement for traditional imputation techniques. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). The exploratory data analysis is performed to gain deeper insights about dataset. The stroke vs no stroke by BMI shows that the most likely to get stroke are the ones whose BMI is nearby 40. The age vs bmi shows that the most likely age to get a stroke is 40 to 60. The

average glucose level should be above 150 to get the brain stroke. The data for no stroke is 4861 while data for stroke is 249. This shows that the model is highly biased towards predicting no stroke. The predictions given by the model are false negative values. To overcome this, SMOTE technique is applied. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class (stroke data). This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

80% of data is used for training while 20% of data is used for testing. Random Forest, Decision Tree, Gradient Boost, SVC, KNN and logistic regression are used to train the model. These 6 classifiers are stacked by using stacking classifiers. Stacked generalization consists in stacking the output of an individual estimator and using a classifier to compute the final prediction. Stacking allows us to use the strength of each individual estimator by using their output as input of a final estimator. Final estimator is also called a meta learner.




The multi-level stacking consists of more than one meta learner and combining them in level. Logistic Regression is used as a final estimator as according to research it is discovered that logistic regression maps the output well from all of the classifiers and helps in avoiding overfitting. K-fold cross validation is used to avoid overfitting too. When a model is trained using all of the data in a single short and give the best performance accuracy. To resist this k-fold cross-validation helps us to build the model is a generalized one. The model is executed with the accuracy of 95%.




To deploy this model, pickle or pkl file is created. Pickle module is a popular format used to serialize and deserialize data types. This format is native to Python, meaning Pickle objects cannot be loaded using any other programming language. It helps to load data in webapp using Flask. The data is then converted into an array and then the data is reshaped again and predicted with the help of a pickle file. HTML form is created and routing is done using @app.route function which helps to route between html and flask. Also template is rendered to fetch the index.html and other html files. Based on the pickle file, the output is given by the Webapp.

## 6. Results & Analysis

The dataset used for the proposed system's performance analysis comprises 5110 rows and 11 attributes, with 80% used for training and 20% used for testing. The 11 attributes are gender, age, hypertension, heart disease, ever_married, residence_type, average glucose level, bmi, smoking status. The dataset is highly imbalanced. The web application is developed using flask. Flask is a web application framework written in python. There is an in-built development server flask.

Figure 5 shows the home page of the proposed system allowing the user to check whether they have a stroke or not.

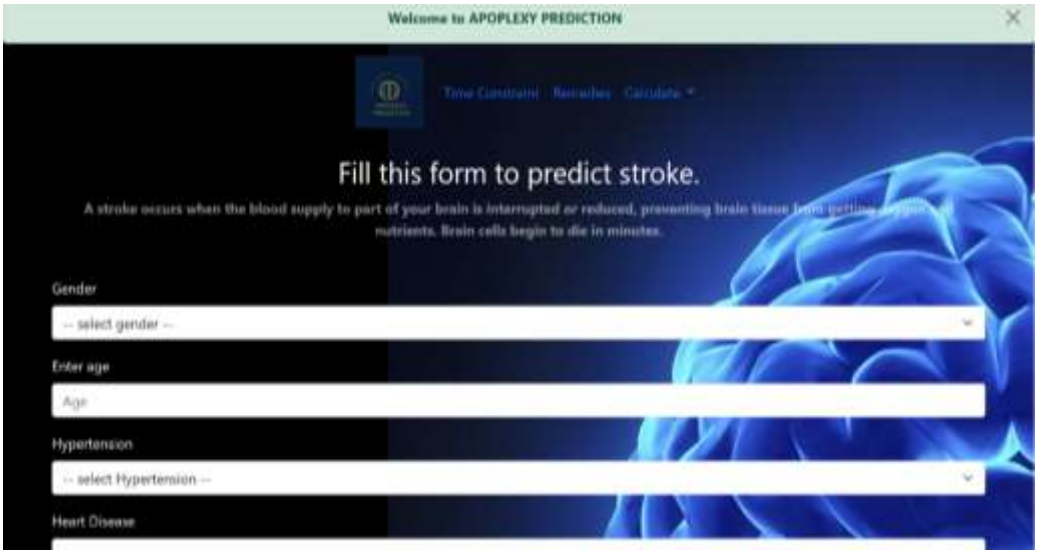



Figure 5: Home Page

Figure 6 shows the result of the brain stroke prediction after taking the input from the user.

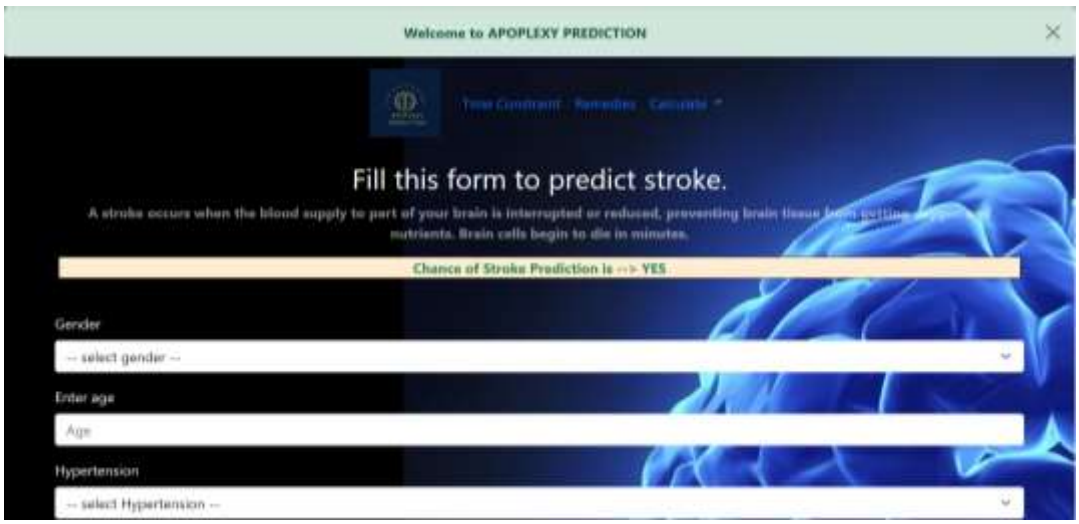



Figure 6: Result Page

Figure 7 shows the time constraint that shows the chance of getting a stroke.


Figure 7: Time Constraint Page

Figure 8 Shows the BMI page that helps to calculate user BMI with the help of user height and weight.

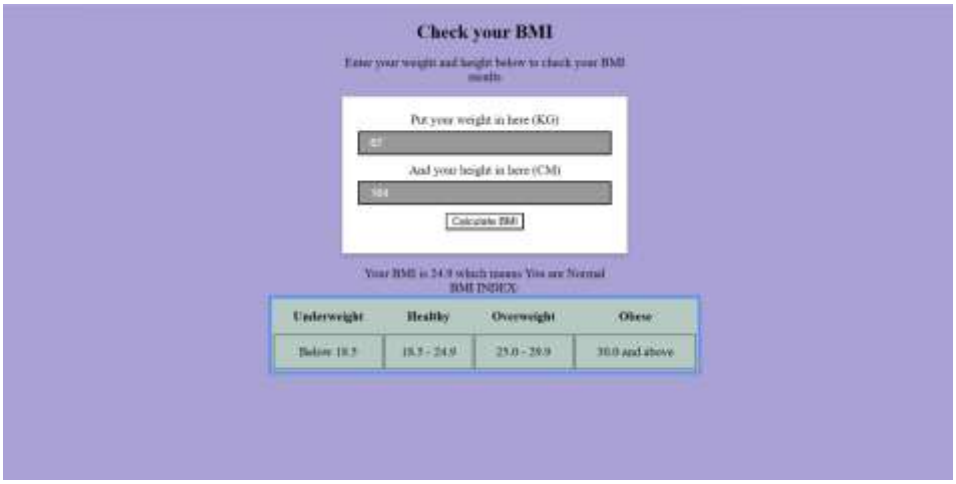



Figure 8:BMI Page

Figure 9 Shows the medicines that are available for brain stroke.

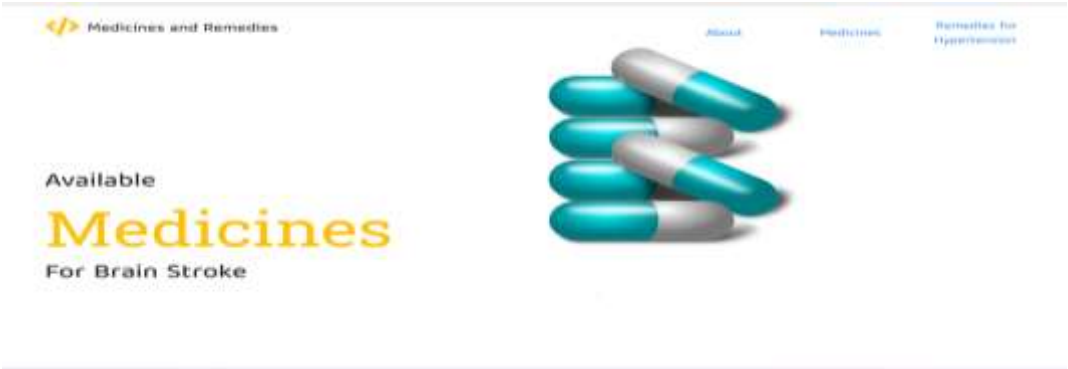



Figure 9: Remedies Page

Figure 10 Shows the hypertension page that helps user to calculate hypertension with the help of systolic and diastolic range.

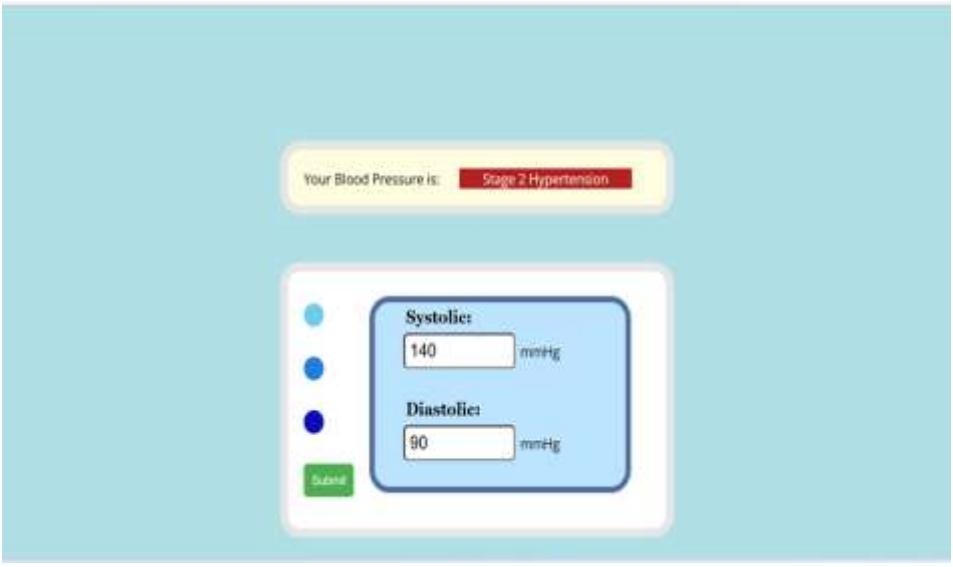



Figure 10: Hypertension Page

Table 4.1 Shows the Accuracy of Each Model and the Accuracy of Proposed Multilevel Stacking Model is 95%

**Table 1 – Accuracy of Models**

| **Model** | **Accuracy** |
|---|---|
| **Decision Tree** | 89% |
| **Random Forest** | 94% |
| **Logistic Regression** | 79% |
| **K-Nearest Neighbor** | 89% |
| **Gradient Boost** | 86% |
| **Support Vector Classifier** | 76% |
| **Stacking** | 94% |
| **Multi-level Stacking** | 95% |

## 7. Conclusion

After the literature survey, there are various pros and cons of different research papers and thus, proposed a system that helps to predict brain strokes in a cost effective and efficient way by taking few inputs from the user side and predicting accurate results with the help of trained Machine Learning algorithms. Thus, the Brain Stroke Prediction system has been implemented using the given 6 Machine Learning algorithm with the help of Multilevel Stacking The system is therefore designed providing simple yet efficient User Interface design with an empathetic approach towards their users and patients. The system has a potential for future scope which could lead to better results and a better user experience. This will help the user to save their valuable time and will help them to take appropriate measures based on the results provided.

**REFERENCES**


[1]. Chirag Rana, Nikita Chitre, Bhargavi Poyekar, Pramod Bide, “Stroke Prediction using Smote Tomek and Neural Network”, IEEE, 2022

[2]. Samaa Ahmed Mostafa, Doaa Saad Elzanfaly, Ahmed El Sayed Yakoub, “Machine Learning Models for Predicting Brain Strokes”, IEEE, 2022

[3]. Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, “Performance Analysis of Machine Learning Approaches in Stroke Prediction”, ResearchGate, 2021

[4]. Kunder Akash, H.N Shashank, “Prediction of Stroke using Machine Learning”, ResearchGate, 2020

[5]. Elias Dritsas, Maria Trigka “Stroke Risk Prediction with Machine Learning Techniques”, MDPI, 2022

[6]. Jose Alberto Tavares, “Stroke prediction through Data Science and Machine Learning Algorithms”, Techtarget, 2021

[7]. Tahia Tazin , Md Nur Alam , Nahian Nakiba Dola , “Stroke Disease Detection and Prediction Using Robust Learning Approaches”, NIH, 2021

[8]. Jiayuan Ling, Gangmin Li, “A two-level stacking model for detecting abnormal users in Wechat activities”, IEEE, 2020

[9]. Shailendra Singh; Abdulsalam Yassine; Rachid Benlamri, “Internet of Energy: Ensemble Learning through Multilevel Stacking for Load Forecasting”, IEEE, 2020

[10]. Abd Mizwar A. Rahim, Andi Sunyoto, Muhammad Rudyanto Arief, “Stroke Prediction using Machine Learning Method with Extreme Gradient Boosting Algorithm”, Metrik, 2022

[11]. Muath A. Obaidat, Alex Alexandrou, Samantha Sanacore, “Machine Learning Stacking Ensemble Model for Predicting Heart Attacks”, IEEE, 2022

[12]. Lesi Nnadozie, Daniel Matthias, E.O Bennett, “A Model for Real Estate Price detection using Multilevel Stacked Ensemble”, EAJ, 2022

[13]. Shivani Batra , Rohan Khurana , Mohammad Zubair Khan, “A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records”, NIH, 2022

[14]. Homay Danaei Mehr, Huseyin Polat , “Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques”, Springer, 2021

[15]. Pradeep Kumar Kushwaha, M. Thirunavukkarasan,"GA-SLE: A hybrid algorithm for heart disease prediction using feature selection Mechanism", Research Square, 2022

[16]. Nourhan M. Swelam, Ayman E. Khedr, "Breast Cancer Diagnosis and Prognosis using stacking ensemble technique", JATT, 2022

[17]. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset