

Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models

Submitted in partial fulfillment of the requirements for the degree of

**Bachelor of Technology
in
Computer Science and Engineering**

by

PRASON POUDEL

19BCE2550

MICKEY KUMAR ROUNIYAR

19BCE2520

**Under the guidance of
Prof. Umamaheswari M**

**Scope
VIT, Vellore.**



July, 2023

DECLARATION

I hereby declare that the thesis entitled “Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models” submitted by me, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering* to VIT is a record of bonafide work carried out by me under the supervision of prof. Umamaheswari M.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore
Date : July, 2023



Signature of the Candidate

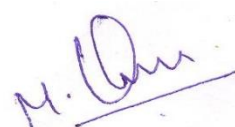
CERTIFICATE

This is to certify that the thesis entitled “Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models” submitted by Prason Poudel 19BCE2550 & Mickey Kumar Rouniyar 19BCE2520, SCOPE, VIT, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering*, is a record of bonafide work carried out by him under my supervision during the period, 01. 03. 2023 to 29.07.2023, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date : July, 2023



Signature of the Guide

Internal Examiner

External Examiner

Dr. Vairamuthu S

B. Tech Computer Science and Engineering

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Umamaheswari M for her invaluable guidance, insightful suggestions, and unwavering support throughout the course of this project. I am also grateful to Vellore Institute of Technology, Vellore for providing me with the necessary resources and facilities to conduct this research. Without their support, this project would not have been possible. Thank you for your invaluable contribution to my Capstone Project.

I would also like to thank the staff and management of Vellore institute of technology, Vellore for their cooperation and support in providing access to the necessary resources and information required for the project. The information and data provided by them were instrumental in conducting the research and analysis.

Thank you all once again for your invaluable support and encouragement throughout this project.

Prason Poudel 19BCE2550

Mickey Kumar Rouniyar 19BCE2520

Executive Summary

The project "Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models" aims to create a website for people to do a self-risk assessment based on various health parameters and machine learning models. The website will take in user inputs related to their health parameters, such as Gender, age, hypertension, heart disease, work type, residence type, bmi, average glucose level, married status and smoking habits and predict the risk of heart stroke.

To accomplish this, we trained various machine learning models, including logistic regression, random forest, SVM, and others. We also created a custom model that takes into account various factors that can affect the risk of heart stroke. We then compared the performance of each model based on their accuracy, sensitivity, specificity, and other metrics. The website works by taking the user inputs, and then passing them through all the trained models. The output from each model is then averaged to determine the overall risk of heart stroke. The website then displays the user's risk level and provides recommendations for reducing their risk.

Overall, the project's objective is to create a user-friendly tool that individuals can use to assess their heart stroke risk and take preventive measures based on their results. The comparative analysis of different machine learning models allows for a more accurate prediction of risk, and the website provides personalized recommendations for each user.

CONTENTS

Page No.

Acknowledgement	i
Executive Summary	ii
Table of Contents	lii
List of Figures	ix
List of Tables	xiv
Abbreviations	xvi
Symbols and Notations	xix
1 INTRODUCTION	12
1.1 Theoretical Background	12
1.2 Motivation	12
1.3 Aim of the Proposed Work	13
1.4 Objective(s) of the Proposed Work	13
2. Literature Survey	14
2.1. Survey of the Existing Models/Work	14
2.2. Summary/Gaps identified in the Survey	14
3. Overview of the Proposed System	15
3.1. Introduction and Related Concepts	15
3.2. Framework, Architecture or Module for the Proposed System	15
3.3. Proposed System Model	16
4. Proposed System Analysis and Design	20
4.1. Introduction	20
4.2. Requirement Analysis	22
4.2.1.Functional Requirements	
4.2.1.1. Product Perspective	
4.2.1.2. Product features	
4.2.1.3. User characteristics	
4.2.1.4. Assumption & Dependencies	
4.2.1.5. Domain Requirements	
4.2.1.6. User Requirements	

4.2.2.Non-Functional Requirements	26
4.2.2.1. Product Requirements	
4.2.2.1.1. Efficiency (in terms of Time and Space)	
4.2.2.1.2. Reliability	
4.2.2.1.3. Portability	
4.2.2.1.4. Usability	
4.2.2.2. Organizational Requirements	27
4.2.2.2.1. Implementation Requirements (in terms of deployment)	
4.2.2.2.2. Engineering Standard Requirements	
4.2.2.3. Operational Requirements (Explain the applicability for your work w.r.to the following operational requirement(s))	
• Economic	
• Environmental	
• Social	
• Political	
• Ethical	
• Health and Safety	
• Sustainability	
• Legality	
• Inspectability	
4.2.3.System Requirements	29
4.2.3.1. H/W Requirements (details about Application Specific Hardware)	
4.2.3.2. S/W Requirements (details about Application Specific Software)	
5. Results and Discussion	30
6. References	40
APPENDIX	

List of Figures

Figure No.	Title	Page No.
1	Architecture for the proposed system	15
2	Architecture of Artificial Neural Network	18
3	Data Distribution	31
4	Outlier Detection	31
5	Important attributes for stroke detection	31
6	Correlation Heatmap of Attributes	32
7	ROC-AUC score and F1 score of ML models	33
8	Accuracy model and loss model of ANN model	34
9	Confusion Matrix of ANN model	34
10	Accuracy model and loss model of CNN model	35
11	Accuracy model and loss model of LSTM model	35
12	Confusion matrix of LSTM model	36
13	Training & Testing accuracy model of Neural network model	36
14	Training and Testing loss model of Neural network model	37
15	Accuracy comparison of all models	37
16	User Interface of the Website	38
17	User input – 1	38
18	Predicting stroke risk based on the average accuracy score	38
19	Heart stroke detection Interface	39
20	User input – 2	39
21	Predicting stroke risk based on the average accuracy score	39
22	Heart stroke not detected interface	39

List of Tables

Table No.	Title	Page No.
1	Accuracy Score of all Models	32

List of Abbreviations

ANN	Artificial neural network
RNN	Recurrent neural network
LSTM	Long short-term memory networks
RF	Random Forest Classifier
DT	Decision Tree
Conv1D	Temporal convolution
KNN	K- Nearest Neighbour
XGBoost	Extreme Gradient Boosting
SVM	Support vector machine
RF	Random Forest
GPU	Graphics Processing Unit
TPU	Tensor Processing Unit
CPU	Central Processing Unit
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
UI	User Interface
CNN	Convolutional Neural Network

Symbols and Notations

%	Percent
~	tilde
*	Convolution Function
Σ	Sigmoid
O	Output Gate

1. INTRODUCTION

1.1.Theoretical Background

Heart stroke is a major cause of death and disability worldwide, and early detection of individuals at risk can help in preventing the onset of stroke. Machine learning (ML) models have shown promising results in predicting the risk of stroke based on various health parameters, such as age, gender, blood pressure, cholesterol levels, and smoking habits.

Logistic regression, random forest, and support vector machines (SVM) are some of the commonly used ML algorithms for predicting stroke risk. Logistic regression is a binary classification algorithm that estimates the probability of an event occurring based on input variables. Random forest is an ensemble method that combines multiple decision trees to improve the accuracy of predictions. SVM is a supervised learning algorithm that identifies the hyperplane that maximally separates data points of different classes. In addition to these standard models, custom ML models can also be developed based on the specific requirements of the project. These models can combine different algorithms or incorporate additional features to improve the accuracy of predictions.

The website built for this project allows users to input their health parameters and receive a risk score for stroke. The average score from all the ML models is used to determine the overall risk level. The website can be a useful tool for individuals to assess their risk of stroke and take proactive measures to prevent the onset of the disease.

1.2.Motivation

The motivation behind the project "Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models" is to help people better understand their risk of experiencing a heart stroke based on their health parameters. Heart stroke is a life-threatening condition, and early detection of its risk factors is crucial for prevention and timely treatment.

By training various machine learning models and building a custom model, this project aims to compare the performance of different models in predicting the risk of heart stroke. The website created as part of the project will provide a user-friendly interface for people to input their health parameters and receive an estimate of their risk of experiencing a heart stroke.

1.3. Aim of the proposed work

The aim of the project "Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models" is to develop a tool that can accurately predict the risk of heart stroke in individuals based on various health parameters. To achieve this, the project involves training various machine learning models such as logistic regression, random forest, and support vector machines (SVM), among others, and comparing their performance to identify the best model for the task.

Additionally, the project involves creating a custom machine learning models and using deep learning models that can improve the accuracy of the predictions. The models are then trained on a combination of features selected based on domain knowledge and data analysis techniques.

Furthermore, the project aims to provide a user-friendly web application that allows individuals to perform self-risk assessments based on their health parameters. The web application will take input from users and provide an average score based on the predictions of all the machine learning models used in the project.

Overall, the project aims to provide an accurate and reliable tool that can help individuals assess their risk of heart stroke and take necessary precautions to prevent it.

1.4. Objective of the proposed work

The objective of the project "Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models" is to develop a system that can accurately predict the risk of heart stroke in an individual based on their health parameters. The project aims to compare the performance of different machine learning models such as logistic regression, random forest, support vector machines (SVM), etc to identify the most accurate and efficient model for predicting heart stroke risk.

The project also involves building neural network models that takes into account additional factors that may contribute to the risk of heart stroke. The custom model will be compared with the other models to determine its effectiveness in predicting heart stroke risk.

Furthermore, the project involves developing a website that allows individuals to assess their own risk of heart stroke by entering their health parameters. The website will use the average score of all the models to detect the risk level and provide personalized recommendations to reduce the risk of heart stroke.

2. LITERATURE SURVEY

2.1. Survey of the existing models/work

- **Logistic Regression** - This is a statistical model used to analyze the relationship between a dependent variable and one or more independent variables. It is commonly used for binary classification problems, such as predicting whether a patient is at risk of a heart stroke or not.
- **Random Forest** - This is an ensemble learning method that combines multiple decision trees to improve the accuracy of predictions. Random forests can handle large datasets with many features and are often used for classification problems.
- **Support Vector Machines (SVM)** - This is a powerful algorithm for classification and regression analysis that works by finding the optimal hyperplane that separates data into different classes. SVM can handle both linear and non-linear classification problems.
- **Neural Networks** - This is a deep learning algorithm that mimics the structure and function of the human brain. Neural networks can handle complex data and can be used for classification, regression, and clustering problems.
- **Decision Trees** - This is a simple and intuitive algorithm that works by splitting data into subsets based on the values of one or more features. Decision trees are often used for classification and regression analysis.

2.2. Summary/gaps identified in the survey

In the survey of machine learning models for predicting heart stroke risk, we have covered some common models such as logistic regression, random forest, SVM, neural networks, and decision trees. We have also highlighted the importance of creating a custom model to tailor the solution to the specific dataset and problem. Additionally, it is important to carefully consider feature selection and preprocessing techniques to ensure that the models are accurately predicting risk based on relevant factors.

Furthermore, it would be beneficial to assess the performance of the models using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). Additionally, it would be useful to perform cross-validation to ensure that the models are not overfitting to the training data. Finally, it would be valuable to validate the models using real-world data and conduct a thorough comparison of the different models to determine which one provides the best performance and accuracy for predicting heart stroke risk.

3. OVERVIEW OF THE PROPOSED SYSTEM

3.1.Introduction and related concepts

The purpose of the project "Predicting Heart Stroke Risk: A Comparative Analysis of Machine Learning Models" is to create a predictive model that can evaluate an individual's risk of having a heart attack based on a variety of health indicators.

Additionally, a website that enables users to self-assess their risk by entering certain health criteria is being built as part of the initiative. The website will compute the average score of all trained models during the prediction process and use it to determine the user's risk of having a heart attack.

Heart strokes are a serious health problem that can result in permanent health issues or even death. Accurate risk prediction of heart stroke can help with early diagnosis and preventative steps to lower the chance of a heart attack.

Machine learning is a potent technique that can be used to precisely forecast the risk of heart attack. The many machine learning models employed in this study will aid in assessing their efficacy and identifying the top model for precisely forecasting the risk of heart attack. The website created as part of this project will offer a user-friendly interface for people to estimate their risk of heart attacks and take the necessary precautions to reduce it.

3.2.Architecture for the proposed system

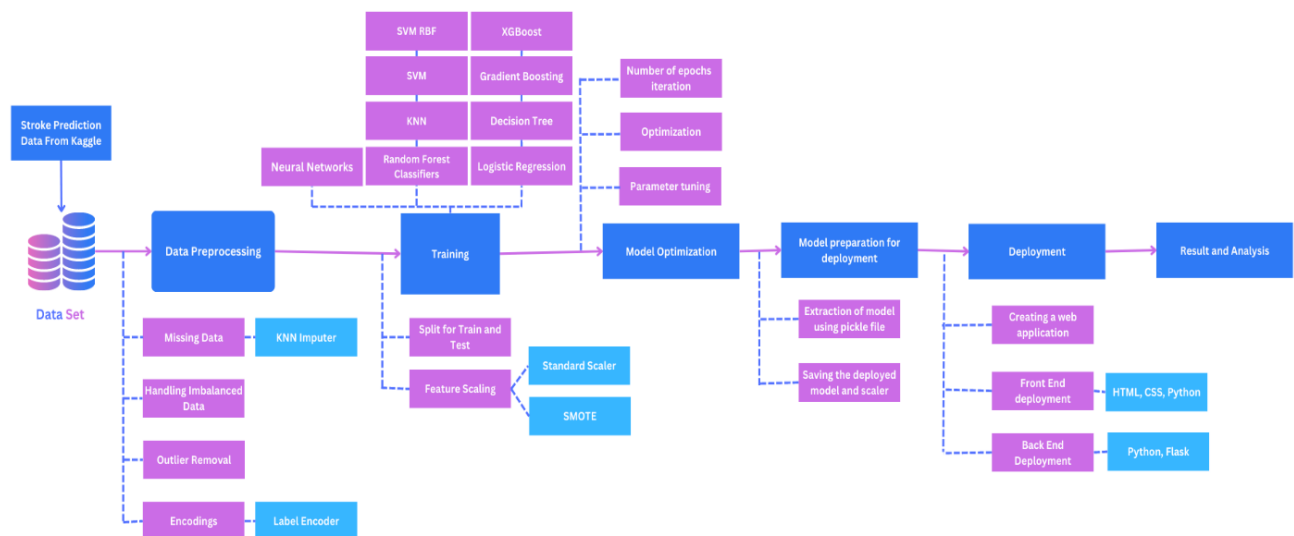


Fig 1: Architecture for the proposed system

The design of the proposed system includes model training, model evaluation, deployment, data gathering, and data preparation. The technology predicts the risk of heart attack using a variety of machine learning techniques and a unique model. Users can perform a self-risk assessment using the system on a website based on a variety of health factors. Based on all of the trained models used to identify the risk, the system produces an average score.

- We import necessary libraries such as pandas, NumPy, matplotlib, scikit-learn, and TensorFlow.
- Collect Dataset which contains features of health parameters.
- Preprocess the data by adding missing data, removing unnecessary columns, removing the outliers by single outlier as a mean of all outliers and converting the health parameters values using label Encoder to binary values 1 and 0.
- Visualize the data by using various plots, such as bar charts, histograms, box plots etc.
- Calculate the correlation between the features and identify the highly correlated features with the target variable.
- Train and evaluate different machine learning models, such as KNN, logistic regression, random forest, decision tree using the preprocessed data.
- Use an ensemble method of these methods and use a voting classifier.
- Train a neural network model using TensorFlow to predict the heart stroke risk based on health parameters.

3.3. Proposed system model

Logistic regression model

For machine learning applications involving binary classification, logistic regression is a powerful tool. By simulating the interaction between a collection of input parameters (features) and a binary outcome variable, it is feasible to roughly predict which class the variable will fall under. A linear combination of input variables is transformed into a probability value between 0 and 1 by the logistic regression equation using a logistic function, often known as the sigmoid function.

$$P(y=1/X) = 1 / (1 + e^{(-z)})$$

- $P(y=1|X)$ is the likelihood that the outcome variable will be 1, given the characteristics of the input X .
- Z is the linear combination of the weighted coefficient-corresponding input feature characteristics.

Random Forest Classifier

A large number of independent decision trees that were each trained separately on a random sample of data make up RFs. The decision trees' outputs are gathered after training, when these trees are built. This algorithm's ultimate forecast is decided by a process known as voting. To choose between the two output classes (in this case, stroke or no stroke), each DT in this procedure must cast a vote. The RF technique selects the class with the highest votes for the final forecast. It can be used for grouping and relapse detection tasks, and it is clear how much importance is placed on each information characteristic.

It is a useful method as well because the default hyperparameters it uses frequently produce clear expectations. Since there aren't many hyperparameters to begin with, it's important to understand them. Although it rarely happens with the arbitrary random forest classifier, overfitting is a well-known issue in machine learning. The classifier won't overfit the model if there are enough trees in the forest.

Decision Tree

Classification and regression difficulties are both addressed by using DT in classification. Additionally, because each input variable has a corresponding output variable, this methodology is a supervised learning model. It appears to be a tree. The data is continuously separated based on a specific parameter in this process. The leaf node and the decision node are the two components of a decision tree. The first node divides the data, and the second node produces the result. DT is easy to comprehend since it replicates the processes a person takes when making a decision in the real world.

Artificial Neural Network (ANN)

The term "artificial neural network" is derived from the biotic neural networks that define the architecture of the human brain.

Similar to the real brain, artificial neural networks have interconnected neurons at different layers of the network. Neurons can also be referred to as nodes.

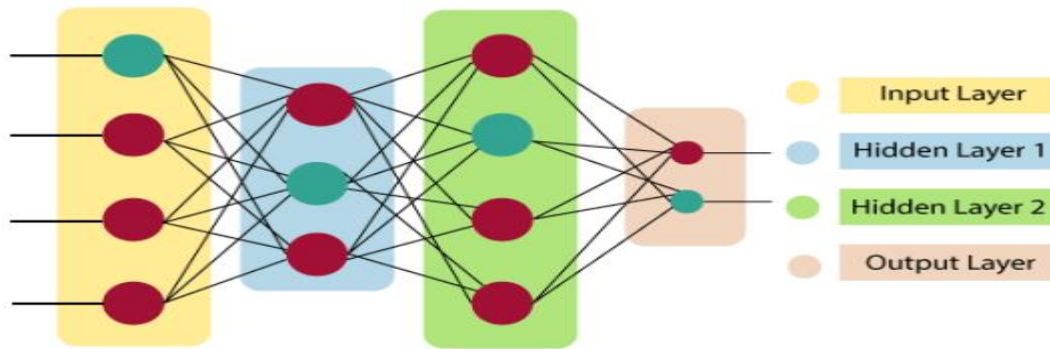


Fig 2: Architecture of Artificial Neural Network

This approach uses input, hidden, and output layers among other layers. The layer's name is input. As suggested by the name, it requires a diversity of input from the programmer. Hidden is the name of the layer. The layer is disguised between the layer input and the layer output. It performs all calculations to check for any buried patterns or structures. The layer's name is output. The input layer is put through a series of transformations using a hidden layer to arrive to the output produced using this layer. When an input is received, an artificial neural network (ANN) computes the input's weighted total while accounting for bias. This computation is stated descent in the form of the transmission function.

XGBoost Classifier

The advanced machine learning technique known as "extreme gradient boosting" (XGBoost) combines the efficiency of gradient boosting with a highly simplified implementation.

To generate predictions, it combines many decision trees. As part of the XGBoost algorithm, the weighted predictions from several decision trees are also incorporated. Each tree learns from the errors of the previous tree. By summing together each of the trees' forecasts, the ultimate forecast is determined. For the purpose of limiting model complexity and avoiding overfitting, regularization terms are also added into XGBoost. The loss function and regularization terms that make up the XGBoost objective function are used to calculate the prediction error of the model. When employing gradient descent to train a model, the target function is improved iteratively.

K-Nearest Neighbor

KNN is a simple, non-parametric supervised learning technique that searches for features in the training set that are similar to one another [38]. It frequently relies on the Euclidian, Manhattan, and Minkowski distance algorithms, which calculate how far apart fresh and old data are. Minkowski, Manhattan, and Euclidian distances can be calculated using the following formulas:

$$\begin{aligned} Euclidean &= \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \\ Manhattan &= \sum_{i=1}^k |x_i - y_i| \\ Minkowski &= \left(\sum_{i=1}^k (|x_i - y_i|^q) \right)^{\frac{1}{q}} \end{aligned}$$

SVM

SVM is a well-known supervised learning method that has long been employed in the medical sector to forecast results and address classification and regression problems. Using the best hyperplane, this classification separates the dataset into two classes. An SVM and kernel are connected linearly in the equation below, where S is the SVM, p_j denotes the patterns in the training set, and q_j denotes the class labels for each pattern.

$$f(p) = \sum_{p_j \in S} \alpha_j q_j K(p_j, p) + b$$

Gradient Boosting:

The Gradient Boosting Classifier builds a collection of weak prediction models typically decision trees that collaborate to produce reliable predictions with the use of boosting and gradient descent methods. Initially, the approach uses the training data to create a flimsy model, like a decision tree. The process then repeatedly develops flawed models while focusing on the shortcomings of the preceding models. Every new model is developed using the residuals (the discrepancies between the actual and anticipated values) of the preceding models. By allocating weights based on the effectiveness of each model, predictions from all the models are merged.

4. PROPOSED SYSTEM ANALYSIS AND DESIGN

4.1.Introduction

Here, the method at result deploys a web application that uses all the pretrained models along with our neural network models to give prediction for the input parameters in real time that is given by the user. In the web application that is deployed using Flask and python for backend, all the predictions are first taken into consideration and then the average of all the models are taken so that our web application is not fully dependent on a single model prediction but on all the models that we have created and deployed.

Our proposed system is divided into the following modules that are listed below:

- Data Collection:

The first step in the system is to collect data from various sources, such as electronic health records, medical journals, and surveys. Here, we have taken the dataset from Kaggle and then we loaded it into our jupyter notebook for our further work. The investigation was conducted using data from the stroke prediction dataset. This dataset has 12 columns and 5110 rows.

- Age: *A person's age is indicated by this property (Numerical Data).*
- Gender: *A person's gender is indicated by this property (Categorical Data).*
- Hypertension: *This characteristic indicates whether or not a person has hypertension (Numerical Data).*
- Work type: *This characteristic describes the person's working environment. (Categorical Data).*
- Residence Type: *The living situation is represented by this attribute (Categorical Data).*
- Heart disease: *The presence or absence of a heart condition is indicated by this feature (Numerical Data)*
- Avg glucose level: *This characteristic refers to the average level of a person's blood sugar (Numerical Data).*
- Body Mass Index (BMI): *It represents the total body mass index of the person that is calculated by the specific ratio of their weight and their height (Numerical Data).*
- Ever married: *This feature indicates whether a person has ever been married (Categorical Data).*

- Smoking status: *This shows the person's smoking history and current update. (Categorical Data)*
- Stroke: *This shows if the person had stroke or not. This is our target feature. (Numerical Data)*

- Data Preprocessing

Initially, we started with exploratory data analysis to correctly understand and visualize our data. In this, we started by visualizing the distribution of each feature and their respected classes under the selected feature such as gender was distributed with values like male and female, smoking type was distributed with values likes formerly smoked, smoking, never smoked and unknown class, work type was distributed with values like government job, private job, self-employed, etc. After this, we checked for missing value which was found in BMI column. The missing value was handled by using the KNN imputer method with the K value set to 30 which took the mean of 30 data near to the missing column. Label encoder was used for encoding of our categorical data.

Some of the insights from our data preprocessing and exploratory data analysis are:

- *Age and target variable weak positive relationship (almost .25).*
- *Average glucose level's mean scores on the target have differences between a person who has a stroke or not. But these differences are small.*
- *BMI does not have any significant relationship with the target variable.*
- *A person with hypertension is almost 3.3 time more likely to get stroke than the ones who don't have hypertension.*
- *Male compare to female is more likely to get stroke, but difference between female and male is very small.*
- *A person with heart disease are 4.07 times more likely to get stroke than the ones who don't have heart disease.*
- *A person is married (or married before) are 5.7 times more likely to get stroke than the ones who don't have marriage history.*
- *Self-employed person has more probability to get stroke than other work type.*
- *Person who lives in rural area slightly has more probability to get stroke than a person who lives in rural area. Difference is small.*
- *It is small difference between who smokes and who does not smoke in regard to probability of getting stroke.*

- Model Training

Various machine learning models are trained on the preprocessed data, including logistic regression, random forest, support vector machine, and neural network models such as ANN, LSTM model and Convolutional 1D model, etc. The models are trained using the training set, and their performance is evaluated using the testing set.

For neural network models, various hyper parameters were tuned to find the optimal curves based on the accuracy and validation data set. For parameters that are used, binary crossentropy function was used for this problem statement as our problem is a binary classification problem.

- Model Evaluation:

The performance of each model is evaluated based on various metrics, such as accuracy, F1 score, confusion matrix and specifically for neural network models, curves were analyzed for accuracy, validation accuracy, loss and validation loss. Based on the number of iterations and analysis through the curve generated, the hyper parameters were adjusted accordingly to get the most optimal and accurate results from these models.

- Deployment:

The selected model is deployed on our own website where users can input their health parameters and get a prediction of their heart stroke risk. During prediction, the system takes the average of all the trained models and provides a risk score based on the average score.

4.2.Requirement analysis

4.2.1. Functional requirements

4.2.1.1. Product perspective

The automated heart stroke risk prediction system based on machine learning is an advanced software tool created to help individual predict heart stroke risk accurately and effectively utilizing various health parameters data. It is meant to be incorporated into current clinical procedures, enhancing the knowledge of pathologists and radiologists. On a technical level, the product is based on reliable machine learning models that have been trained on large, complex and annotated datasets. The key patterns and traits suggestive of heart stroke are captured and interpreted by these models using advanced image processing and feature extraction techniques. In order

to guarantee accuracy and generalizability, the models go through rigorous training and validation processes.

4.2.1.2. Product features

The automated heart stroke risk prediction product based on machine learning using datasets offers a range of features to support accurate and efficient heart stroke risk detection. It provides features for organizing, annotating, and importing datasets to ensure effective data management. The product uses data preprocessing methods such as adding missing data, removing outliers, using label Encoder to replace values by 0 and 1, and data imbalance correction. Model training uses machine learning methods, such as deep learning structures like Logistic Regression, Random Forest Classifier, Decision tree, SVM, XGBoost Classifier and Neural Network. Various criteria are used to evaluate and validate the trained models, and model selection and optimization approaches are used to determine which models perform the best overall. ROC, F1 score, recall, and other evaluation metrics are used in the product's processes for assessing and verifying the trained models. This guarantees the model's functionality and generalizability, which helps to reliably and accurately identify heart stroke risk.

4.2.1.3. User characteristics

- Healthcare Professionals:

Healthcare professionals with expertise in the detection and treatment of heart stroke are the main users of automated heart stroke risk detection systems based on machine learning. This covers the clinical judgments made by medical professionals while interpreting health parameters.

- Technical Proficiency:

To interact with the automated heart stroke risk detection system successfully, users must have a specific level of technical competence. They have to be experienced operating computer systems, navigating software user interfaces, and comprehending the fundamental ideas behind machine learning and health parameters.

- Domain Expertise:

Users should have in-depth knowledge of heart stroke, its pathology, and the specifics of health parameters. For evaluating the system's predictions, validating the results, and making decisions based on the automated analysis, this domain expertise is essential.

- Ability to Validate and Interpret Results:

Rather than replacing healthcare experts, automated heart stroke risk detection systems are created to support them. Users should be able to verify and understand the system's findings in light of the patient's clinical background, symptoms, and other health parameters data. Critical thinking, clinical reasoning, and fusing the automated analysis with their own knowledge are required for this.

Users should have a thorough awareness of the ethical issues related to employing automated methods for heart stroke risk detection. They should follow laws governing patient privacy and confidentiality, use technology impartially and fairly, and engage with patients in a way that fosters openness and confidence. By taking into account these essential user characteristics, automated heart stroke risk detection systems can be developed and deployed to specifically suit the requirements of medical professionals, promote teamwork, and enhance the overall diagnostic procedure.

4.2.1.4. Assumption And Dependencies

Several crucial presumptions must be true in order for machine learning-based automated heart stroke risk detection to work. In the beginning, it is assumed that a sufficient and representative dataset is available, one that correctly reflects the variety of heart stroke cases. To guarantee that the models are capable of generalizing adequately to varied circumstances, this dataset includes a variety of health parameters data. For the purpose of extracting significant characteristics and building precise machine learning models, the quality and dependability of the health parameters datasets are essential. It is further considered that the characteristics taken from the health parameters dataset are crucial and indicative of the heart stroke risk. It is anticipated that the chosen feature extraction algorithms would successfully capture

the distinctive qualities of health parameters and identify them from normal health parameters. Finally, it is believed that the annotation and labeling of the health parameters in the dataset are consistent and properly represent the truth in supervised learning settings. The machine learning models are trained using exact labels thanks to the precise and trustworthy annotations provided by expert pathologists.

Developing automated heart stroke risk detection systems based on machine learning requires close cooperation with medical researchers. Their knowledge and input are crucial for confirming the system's effectiveness and assuring compliance with clinical procedures. Large datasets, sophisticated computations, and machine learning model optimization all require access to sufficient computational resources, such as high-performance computing infrastructure and potent GPUs. The credibility, efficiency, and practical application of automated heart stroke risk detection systems are boosted by these resources.

4.2.1.5. Domain Requirements

When addressing various health parameters types, values, and range, the system should demonstrate robustness to variety. When processing vast amounts of health parameters data, it should operate effectively and scale. Additionally, the system should be interpretable and explicable so that medical practitioners may comprehend the variables influencing its forecasts. In order to ensure the secure processing of patient data, privacy and security are essential. Last but not least, user-friendliness and efficient utilization within clinical settings require seamless interaction with the clinical process.

4.2.1.6. User Requirements

High accuracy in detecting heart stroke risk, usability through an intuitive interface integrated into clinical workflows, interpretability to understand the system's decisions, alignment with accepted clinical practices, scalability to handle large datasets, and adherence to privacy and security standards are among the user requirements for an automated heart stroke risk detection system based on machine learning. By fulfilling these criteria, the system will be able to assist medical professionals in the diagnosis and treatment of heart stroke, ultimately leading to improved patient care and outcomes.

4.2.2. Non-Functional Requirements

4.2.2.1. Product Requirements

4.2.2.1.1. Efficiency

- $O(n)$, where n is the total number of rows in the dataset, is the reading time for dataset.
- Depending on the complexity of the procedures carried out, data preprocessing and visualization take either $O(n)$ or $O(n^2)$ time.
- Training and evaluating machine learning models requires training and evaluating several models, hence the time required is $O(n)$.

4.2.2.1.2. Reliability

The method is very reliable since it constantly provides correct predictions while reducing false positives and false negatives. It is strong and adaptable to changes in health parameters, their values, and heart stroke symptoms. The system's capacity to function consistently across multiple datasets and contexts, indicating generalization capabilities.

4.2.2.1.3. Portability

To be easily browsed and used across diverse healthcare settings and platforms, the system must be portable. Healthcare professionals can access and use the application anywhere, regardless of the specific technical infrastructure that may be in place. Computers, laptops, and mobile devices with a range of operating systems and hardware should all function with the system. In order to ensure seamless integration and effective adoption, the system should also be able to interact with currently used healthcare systems and practices. By putting mobility first, the automated heart stroke risk detection becomes more usable and adaptable, enabling healthcare providers to use it in a range of clinical settings and improve patient care.

4.2.2.1.4. Usability

The system is intended to be user-friendly and intuitive, allowing healthcare practitioners to quickly browse and engage with its features. A simple and well-designed user interface makes it easier to input health parameters value, retrieve forecasts, and access important details. Furthermore, the technology

seamlessly integrates into existing clinical operations, assuring little disturbance and maximum productivity.

4.2.2.2. Organizational Requirements

4.2.2.2.1. Implementation Requirements

- **Hardware Infrastructure:** Ensure that the automated heart stroke risk detection system can be deployed with the support of an appropriate hardware infrastructure. This could comprise networking hardware, servers, storage devices, GPUs or TPUs (Tensor Processing Units) for accelerated computations.
- **Software Frameworks and Libraries:** We'll need to ensure that the following libraries must be present in the user system such as TensorFlow, or scikit-learn. Make sure that the frameworks are compatible with the hardware infrastructure and can make use of frameworks that offer deployment optimizations.

4.2.2.2.2. Engineering Standard Requirements

Implement best practices for machine learning model construction, such as appropriate model selection, feature engineering, and hyperparameter tuning. Model construction and Validation. Utilize benchmark datasets and relevant evaluation measures to validate the models. The model development process should be documented, along with the algorithms employed, version control, and performance testing.

- **Preparing Training Data:**

Prepare the training data in accordance with legal and moral standards. In order to safeguard privacy, make sure that patient data is properly anonymized and deidentified. Use data preprocessing methods to improve the quality and diversity of the training dataset, such as normalization, augmentation, or noise reduction.

- Performance measures and Validation Protocols:

Select the best performance measures to evaluate the automated heart stroke risk detection system's accuracy, sensitivity, specificity, precision, and recall. To enable fair evaluation and comparison of various algorithms and models, establish validation processes.

4.2.2.3. Operational Requirements

- Economic: The Automated heart stroke risk detection System would incur cost based on the requirements which includes hardware, software deployment, server costs and maintenance in starting scale.
- Environmental: It is implemented on Computer Systems that do not require higher specs therefore poses no environmental threat.
- Social: It will create a sense of reliability towards doctors and medical practitioners as it helps them yield better results and treatment. This can be used in all forms including personal self-analysis, health institutes to take predictions and act accordingly and be more aware in long run as well as doctors for checkups leading to more reliable treatments.
- Political: The system does not conform to any political bias as it is for the betterment of patients and is related to health which does not obstruct to any form of political disagreement.
- Ethical: Our system makes sure that data privacy laws are strictly followed, and that patient data confidentiality is upheld. Obtain the patient's informed consent before collecting and using their medical information for heart stroke risk detection.
- Health And Safety: We take responsibility for the accuracy and reliability of the automated system but the diagnosis done on the predicted result is solely based on the doctor that is being consulted by the patient. Also, the system developed is just used for predicting the future outcome and take precautions beforehand.

- Sustainability: This system does not require large amount of power and can run on very less energy which means the system can sustain for a long period of time. In future, is scalability is concerned for this, a greater number of models can be deployed in this for taking predictions which make require efficient systems for the model to work with, and application maintenance if expanded can incur more expenses based on the scale it is being expanded to.
- Legality: The dataset used by our automated system is freely available to everyone so it is legally sourced. However, all the patient's data should be strictly confidential if to be used by health institutions and other related professions for diagnosis.

4.2.3. System Requirements

4.2.3.1. Hardware Requirements

- Processor: Minimum i5 Dual Core
- Ethernet connection (LAN) OR Wi-Fi
- Hard Drive: Recommended 100 GB or more
- Memory (RAM): Minimum 8 GB

4.2.3.2. Software Requirements

- Python
- Anaconda
- Jupyter Notebook
- TensorFlow
- Flask

5. RESULT AND DISCUSSION

We initially started our project with a very deep data analysis for our dataset to fully understand our data and its distribution and make changes in specific part from the analysis done in the exploratory data analysis section and data processing part. The performance of each of the models was evaluated using the F1 score, ROC score and accuracy score. Also, for the neural network that is used in our project, the accuracy and loss curves were analyzed for results and confusion matrix to see the percentage and distribution of our data lying in positives and negatives predicting the probability of stroke.

We aimed to develop a website that uses various machine learning models to predict heart stroke risk based on several health parameters. The project involved training multiple machine learning models, including logistic regression, random forest, support vector machine (SVM), and a custom model. The models were then used to create a website where users could input their health parameters and receive a prediction of their heart stroke risk.

The dataset used in the project contained information on various health parameters such as age, gender, blood pressure, and cholesterol levels. The dataset was split into training and testing sets to evaluate the performance of each machine learning model. Performance was evaluated using metrics such as accuracy, precision, recall, and F1 score.

INSIGHTS FROM DATA ANALYSIS

- *Average glucose level's mean scores on the target have differences between a person who has a stroke or not. But these differences are small.*
- *BMI does not have any significant relationship with the target variable.*
- *A person with hypertension is almost 3.3 times more likely to get a stroke than the ones who don't have hypertension.*
- *Male compared to female are more likely to get strokes, but the difference between female and male is very small.*
- *A person with heart disease is 4.07 times more likely to get a stroke than the ones who don't have heart disease.*
- *A person who is married (or married before) is 5.7 times more likely to get a stroke than the ones who don't have a marriage history.*
- *Self-employed person has more probability to get strokes than other work types.*
- *Person who lives in a rural area slightly has a higher probability of getting a stroke*

than a person who lives in a rural area. Difference is small. There is a small difference between who smokes and who does not smoke in regard to the probability of getting a stroke.

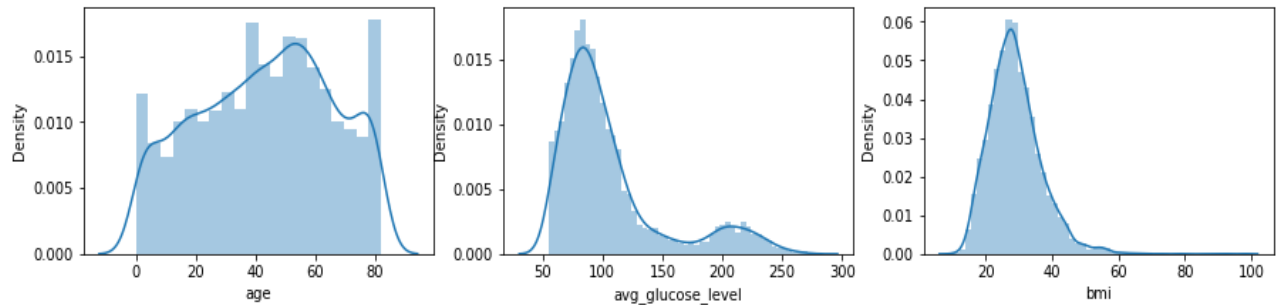


Fig 3: Data Distribution

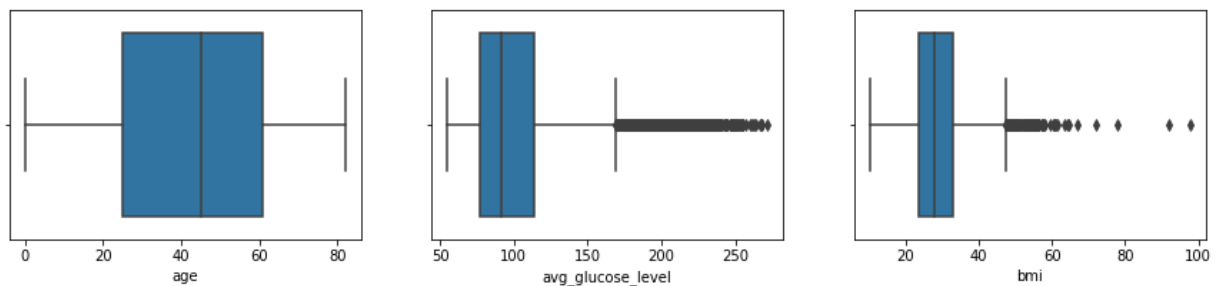


Fig 4: Outlier Detection

After, this we found out that age, BMI and average glucose level were important attributed as their correlation was higher in terms of stroke than any other features which was also indicated in the papers that we referred.

	age	avg_glucose_level	bmi	stroke
age	1.000000	0.238318	0.333244	0.245253
avg_glucose_level	0.238318	1.000000	0.175672	0.131991
bmi	0.333244	0.175672	1.000000	0.042341
stroke	0.245253	0.131991	0.042341	1.000000

Fig 5: Important attribute for stroke detection

- When age increases, also the mean score on the stroke also increases.
- Average glucose level's mean scores have differences between a person who has a stroke or not.

- BMI mean scores are close to each other.
- Correlations with the target variable are very small.
- Among all numerical features, age is the most dominant feature with a correlation of 0.24

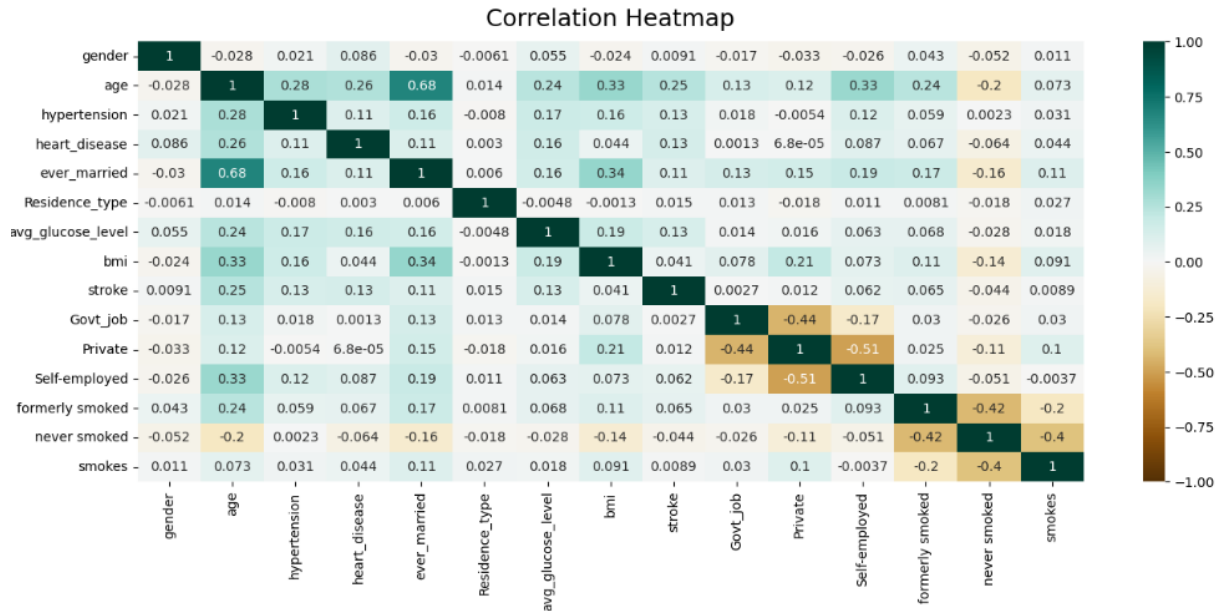


Fig 6: Correlation Heatmap of Attributes

The results of the project showed that all of the machine learning models performed reasonably well in predicting heart stroke risk. Here, after analyzing the results, we found that the neural network models developed for the project performed better and more accurate than the other models. Out of all the pretrained models that were trained and tested in our project, we found out that the logistic regression model and SVM model acquired high level of accuracy out of all the pretrained models that were trained in this project. The accuracy levels of the models are showed below:

MODELS	ACCURACY
LOGISTIC REGRESSION	77.191%
K-NEAREST NEIGHBOUR	60.872%
DECISION TREE	59.193%
SVM (Linear Kernel)	76.73%
SVM (RBF Kernel)	69.043%
RANDOM FOREST	56.405%

GRADIENT BOOSTING	68.39%
XGBOOST	56.611%
ANN	84.44%
LSTM	87%
CONVOLUTIONAL 1D	83.27%

Table 1: Accuracy score of all Models

For the pretrained models, the roc score and f1 score was found to be as follows:

```

Model Performance
-----
Logistic Regression Roc Auc Score: 77.191%
                   F1-Score: 0.23599
K-Nearest Neighbors Roc Auc Score: 60.872%
                   F1-Score: 0.16740
Decision Tree       Roc Auc Score: 59.193%
                   F1-Score: 0.18978
Support Vector Machine (Linear Kernel) Roc Auc Score: 76.728%
                                       F1-Score: 0.22989
Support Vector Machine (RBF Kernel)   Roc Auc Score: 69.043%
                                       F1-Score: 0.20478
Random Forest        Roc Auc Score: 56.405%
                   F1-Score: 0.17978
Gradient Boosting    Roc Auc Score: 68.393%
                   F1-Score: 0.26230
XGBoost              Roc Auc Score: 56.611%
                   F1-Score: 0.18824

```

Fig 7: ROC-AUC score and F1 score of ML models

Now, after we evaluated our pretrained models, we analyzed our neural network model starting with our custom neural network model. In this, we achieved the accuracy on training data set of 92.2% and accuracy on testing data set was found out to be as 82%. Though this had good results, the validation loss was found out to be hiking and was greater in this case which was resolved in other models that we trained and tested.

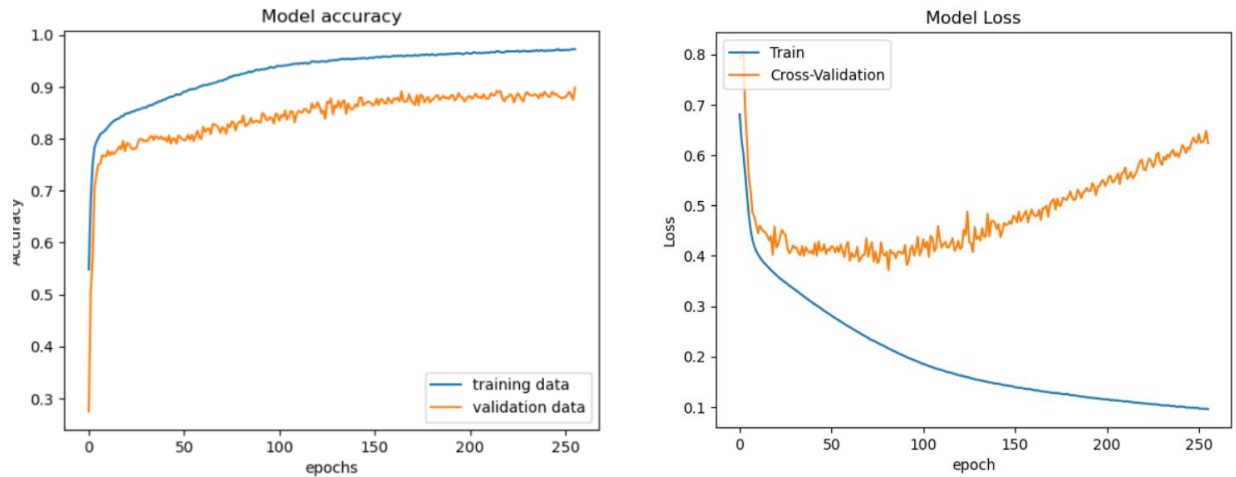


Fig 8: Accuracy model and loss model of ANN model

Here, the confusion matrix generated was:

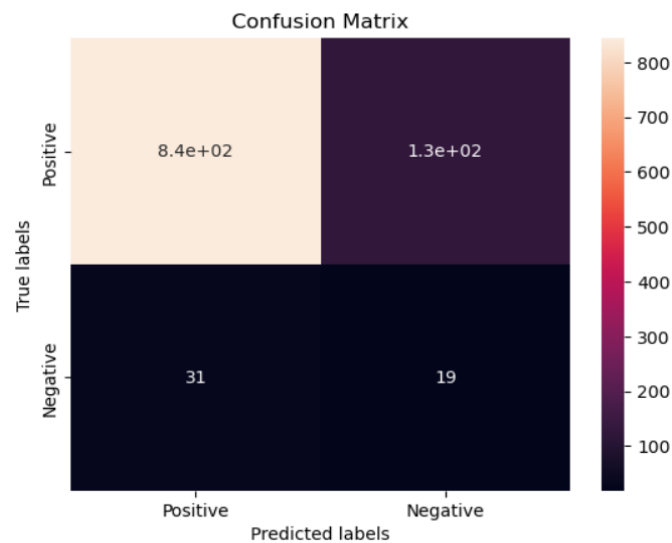


Fig 9: Confusion Matrix of ANN model

Here, we found out that our true positives were in greater frequency but our confusion matrix predicted that nearly 140 of the labels were predicted positive which was originally negative. Though our model received an accuracy of 84.44%, there was scope for improvement which was done by the use of other models.

Upon training and testing of the convolutional 1d model, the training and testing accuracy was found out to be 93.38% and 83.27% respectively. This was a slight improvement than our ANN model that we created and the loss curves generated here was better than the curves that we found in our ANN model.

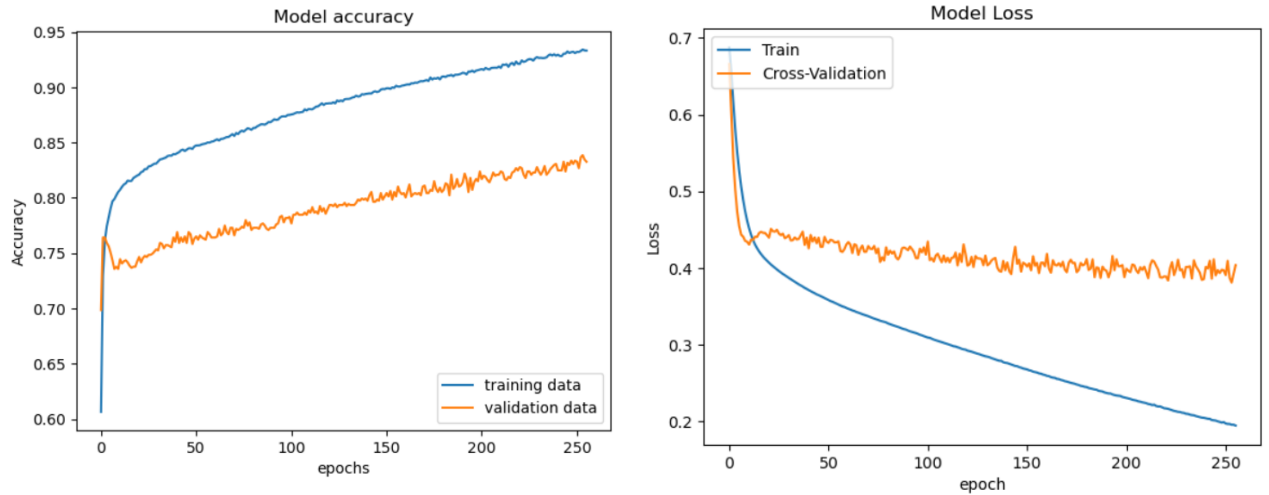


Fig 10: Accuracy model and loss model of CNN model

Finally, we trained and tested our LSTM model by changing the hyper parameters to find the most optimal result. In our LSTM model, we acquired an accuracy of training and testing as 97.67% and 87% respectively. This model was far more accurate than all the neural and pretrained models that was used in this project. The validation and training accuracy and loss curves were found to be optimal and better than any other curves we generated in other neural network models.

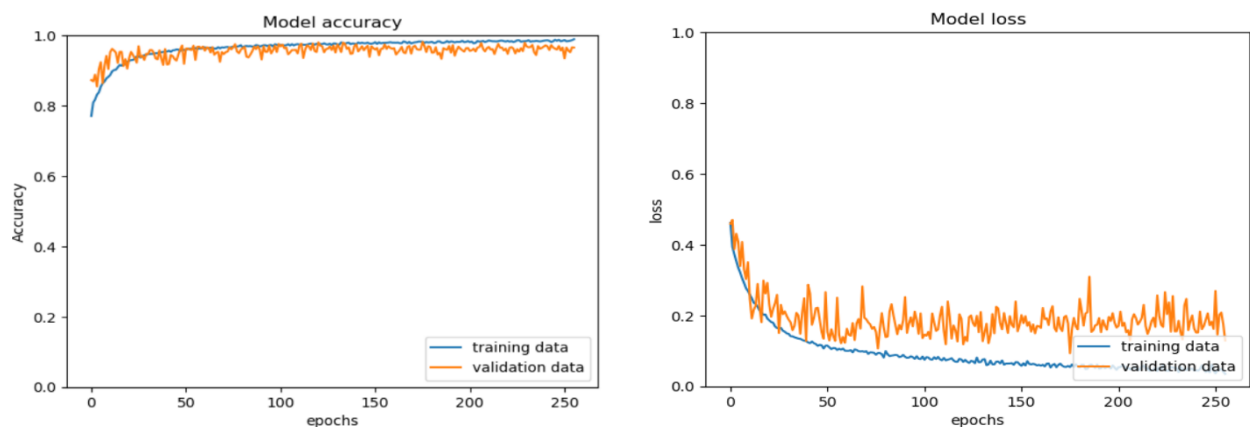


Fig 11: Accuracy model and loss model of LSTM model

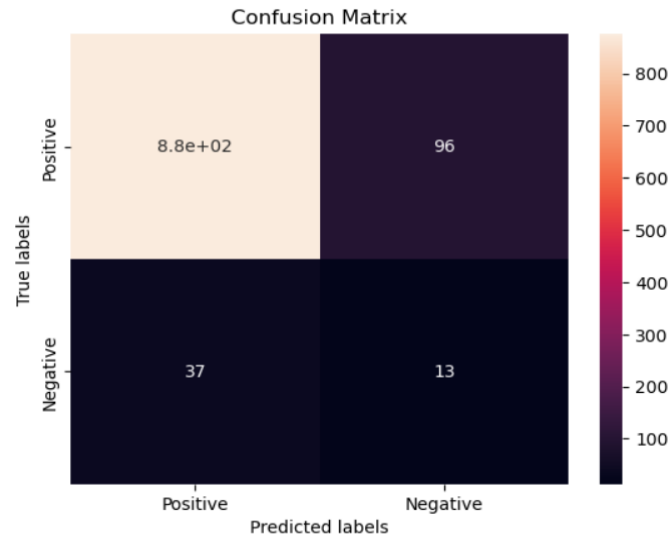


Fig 12: Confusion matrix of LSTM model

Here, as we can see the confusion matrix generated by this model, we can efficiently conclude that the number of false positives and false negatives are very less as compared to the other neural network models and we tested in this project. This indicated that the use of LSTM model was by far more accurate for this problem and accurate.

After analyzing the results of all the neural network models that were used in this project, we can see that in terms of accuracy on training and validation dataset, LSTM outperformed all the other models that were used.

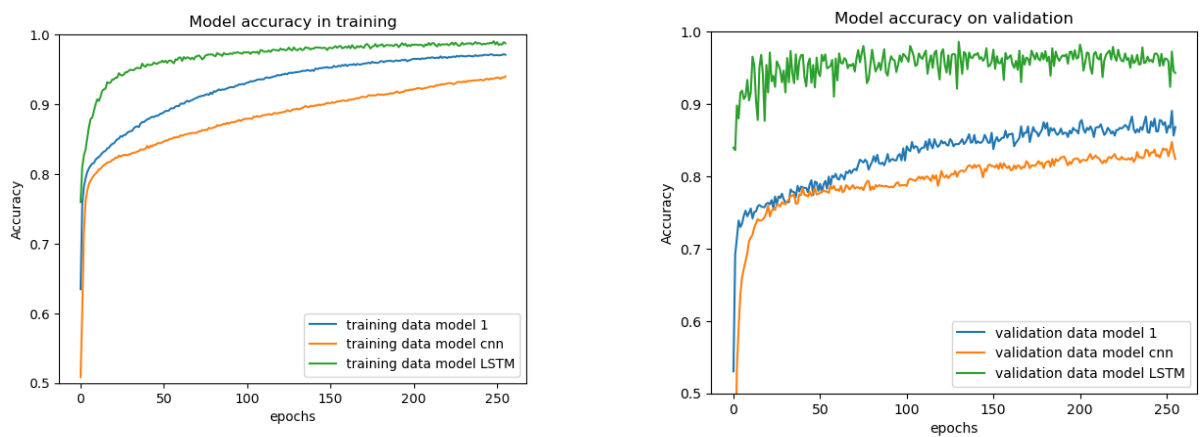


Fig13: Training & Testing accuracy model of ANN, CNN and LSTM model

Similar results were found in training and validation loss.

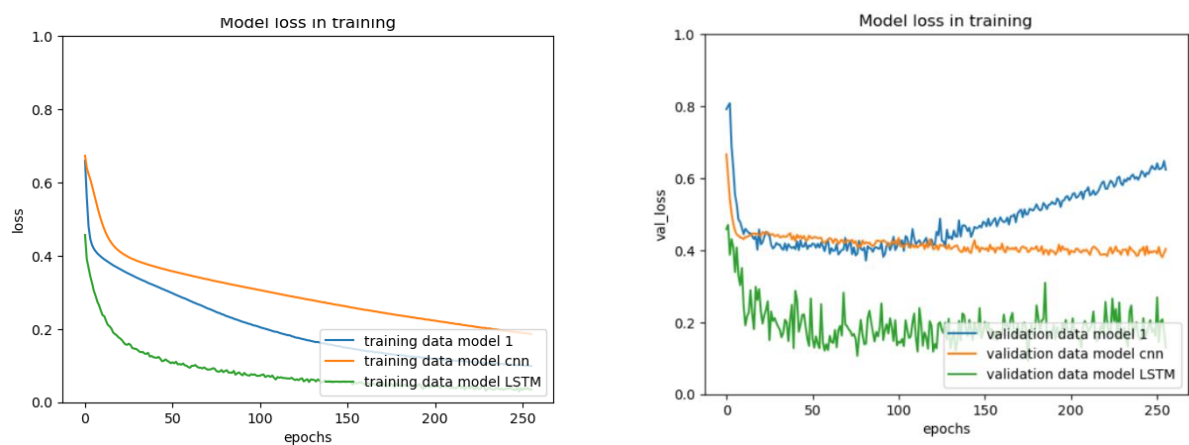


Fig 14: Training and Testing loss model of ANN, CNN and LSTM model

In this loss curves that are plotted above, we can see that in the validation dataset, the loss of our ANN model was is increasing order which was rectified by other models such as Conv1D and LSTM. Out of all the models, we can visualize that the LSTM losses were significantly lower that the other models which showed that the use of LSTM can help in predicting more accurate results which interacting with our web application.

Here, the accuracy of all the models used in this project are visualized showing that the neural network model outperformed all the other pretrained models in this project. Out of the pretrained models, Logistic regression and SVM models showed higher level of accuracy that other models whereas in neural network models that were used, LSTM showed more accurate and promising results in predicting the stroke based on the parameters that were used.

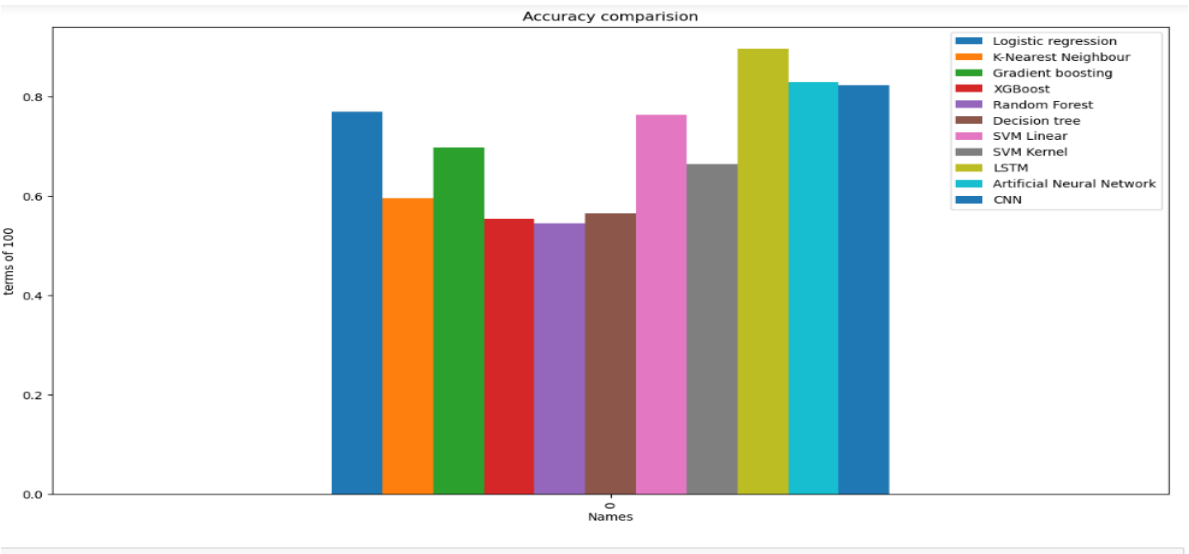


Fig15: Accuracy comparison of all models

The website developed for the project allows users to input their health parameters and receive a prediction of their heart stroke risk. The website uses an ensemble approach, taking the average of all the models' predictions to arrive at the final prediction. This approach is expected to provide a more accurate prediction than any individual model.

Heart Stroke prediction
ANALYSIS OF HEALTH PARAMETERS

Heart Stroke prediction

Personal Details

Select Gender	Age
Hypertension	Heart Disease
Ever married	Work type
Residence type	Avg_glucose_level
Bmi	Smoking Status

Submit

Fig 16: User Interface of the Website

CASE 1:

```
result() > if finalvote < 0.5
app x
[[ 1. 67. 1. 1. 0. 1. 1. 226.98 36. 1. ]]
[[ 1.19094215 1.0486766 3.09204197 4.07631933 -1.39294213 -0.87483256
0.99487489 2.73362938 0.92496595 -0.56365763]]
```

Fig 17: User input - 1

```
app x
1/1 [=====] - 0s 70ms/step
1/1 [=====] - 0s 116ms/step
1/1 [=====] - 0s 452ms/step
[1] [1] [0] [1] [1] [1] [[1.3307773e-06]] [[0.87221986]] [[0.00016003]]
[[0.6524868]]
```

Fig18: Predicting stroke risk based on the average accuracy score

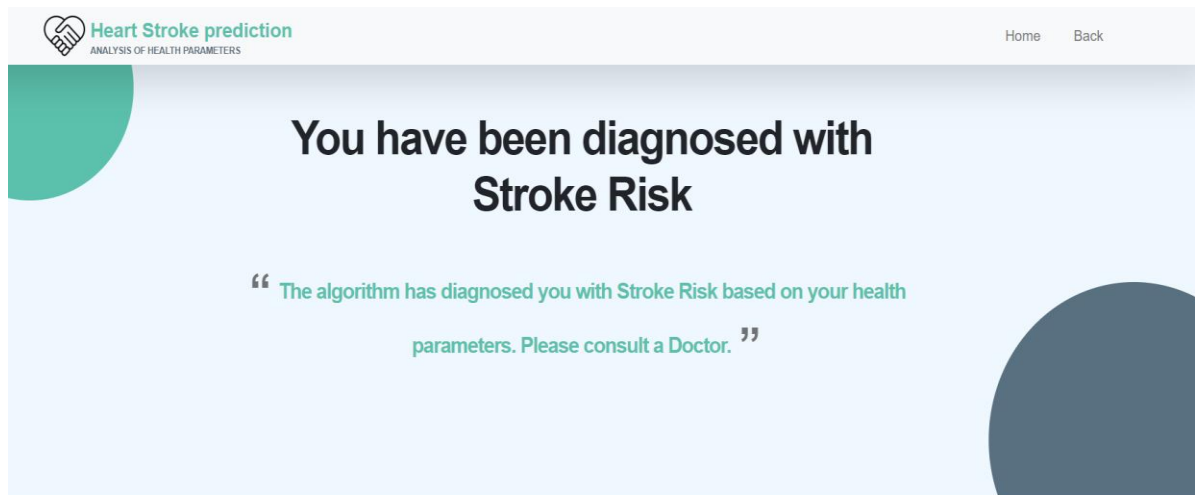


Fig 19: Heart stroke detection Interface

CASE 2:

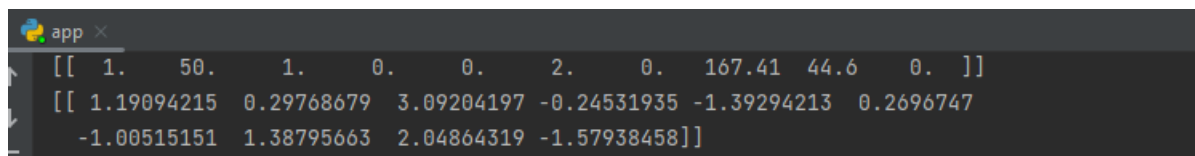


Fig 20: User input - 2

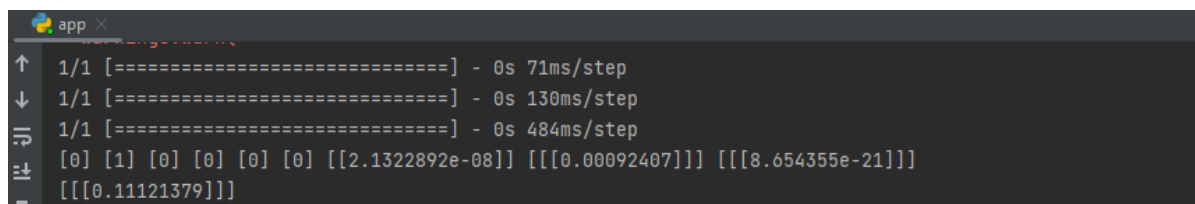


Fig 21: Predicting stroke risk based on the average accuracy score

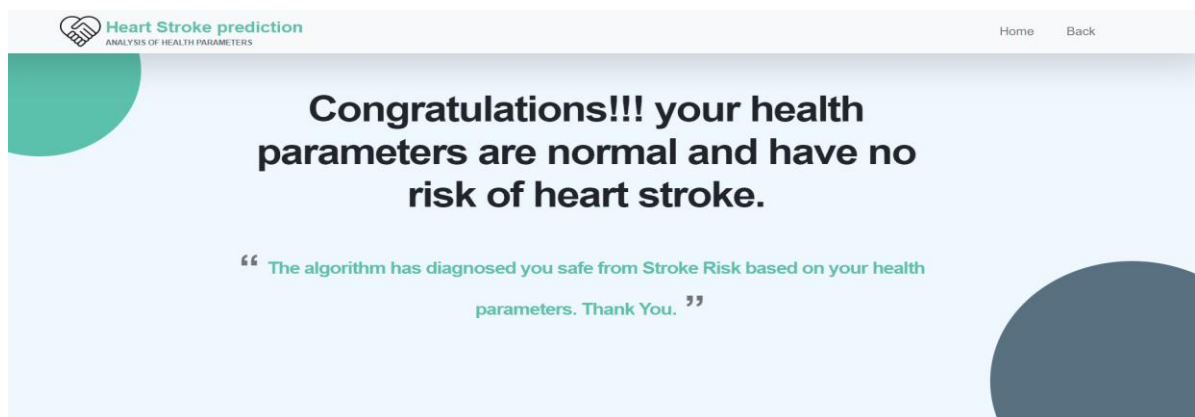


Fig 22: Heart stroke not detected interface

Overall, the project shows that machine learning models can be effectively used to predict heart stroke risk based on various health parameters. The website developed for the project provides a convenient and accessible tool for individuals to assess their own risk of heart stroke.

6. REFERENCES

- [1]. "Stroke Risk Prediction with Machine Learning Techniques", Elias Dritsas, Maria Trigka, 2022
- [2]. "Heart Attack Prediction Using Machine Learning Algorithms ", Manjula P, Aravind U R, 2022
- [3]. "Stroke Disease Detection and Prediction Using Robust Learning Approaches", Tahia Tazin , Md Nur Alam, 2022
- [4]. "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction ", Hassan Dawood, Muhammad Waqar 2021
- [5]. "A predictive analytics approach for stroke prediction using machine learning and neural networks ", Soumyabrata Dev, Hewei Wang, 2022
- [6]. "Heart Stroke Prediction Using Machine Learning ", Syed Shareefunnisa, S N Lakshmi Malluvalasa, 2022
- [7]. "Stroke Prediction Using Machine Learning Classification Algorithm ", Michael Wiryaseputra, 2022
- [8]. "Performance Analysis of Machine Learning Approaches in Stroke Prediction ", Minhaz uddin Emon, Maria Sultana Keya, 2021
- [9]. "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach ", Nitish Biswas, Khandaker Mohammad Mohi Uddin, 2021
- [10]. "A Study of Features Affecting on Stroke Prediction Using Machine Learning" , Panida songram, Chatklaw jareanpon, 2022