Enhancing Stroke Prediction Dataset Performance Through Dataset Merging and Analyzing Important Feature

BY

Rabeya Sultana Riju
ID: 201-15-3062

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Taslima Ferdaus Shuva

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Md Assaduzzaman

Lecturer (Senior Scale)

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

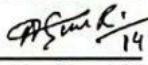
This Project titled "Enhancing stroke prediction dataset performance Through dataset merging and analyzing important feature", submitted by **Rabeya Sultana Riju** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14 July, 2024.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan (AAK) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Aynul Hasan Nahid (AHN) — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Abu Sayad Md. Mostafizur Rahaman (ASMR) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

©Daffodil International University i

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Taslima Ferdaus Shuva, Assistant Professor, Department of CSE, Daffodil International University.** We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma

Supervised by:

Taslima Ferdaus Shuva

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised by:

MR.MD Assaduzzaman

Lecturer (Senior Scale)

Department of CSE

Daffodil International University

Submitted by:

Rabeya Sultana Riju

ID: -201-15-3062

Department of CSE

©Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

I are grateful and wish our profound indebtedness to **Taslima Ferdaus Shuva, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Field name*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori**, Head, Department of CSE, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Predictive modeling plays a vital role in stroke prediction, enabling timely intervention in healthcare. It is inevitable that any form of prediction especially in the prediction of stroke would require some form of predictive modeling that empowers timely healthcare intervention. This work can be used for proposing evaluation criteria of the stroke prediction model concerning the dataset size and feature influence. Two datasets were used: the first one with 5000 data points, the program had a 90% accuracy while the second with 1500 data points had 50% accuracy. Such transitions merged increased the datasets to a combined percentage of 60% to the other's benefit, proving the importance of diverse datasets. Other measures pointing to the exploration of the feature importance analysis include Average Glucose Level and Body Mass Index (BMI) that are important in the accurate prediction of strokes. Therefore, these results present task, data set attributes, and the importance of features as crucial factors to be taken into account to build stable models. The nature of change in the difference in model accuracy between two datasets shows that large samples produce higher model accuracy. Future studies are likely to reveal improved models for different datasets and different subpopulations in order to ultimately enhance process accuracy and consequently stroke predictability as well as its effects on its patients. Understanding these factors can help enrich such determinants among the health-care providers which in turn would lead to improved ways of intervention and management of the risks associated with stroke. As such, this study emphasis the possibilities of predictive modeling to enhance the efficiency and effectiveness of the stroke prevention and care by providing accurate and reliable predictions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The global burden of stroke, a devastating neurological event, remains a considerable concern for public health around the world, with various implications for patients and healthcare systems. In particular, the outcomes of a stroke are not limited to the risk of death but manifest in sustained disability and deterioration of the quality of life for those who live. As such, the need for stroke prediction has become more pressing than ever, and predictive modeling has evolved into an integral part of modern healthcare. Thus, it is hoped that it can assist in identifying the condition before it exacerbates. Nevertheless, the outcome of predictive models is determined by a plethora of factors, the main of which is the characteristics of the datasets that are used to build them. Dataset size, diversity, and quality are the major issues that can affect model accuracy and generalizability, therefore, the careful selection and creation of a dataset is the key point in modeling.

This research aims at a broad survey of the complex dynamics that regulate stroke prediction through the perspective of predictive modeling.[1] We use two different datasets, each being characterized by its own demographic and clinical profiles, to find the finer details in the performance of the models among different data landscapes. Through the application of the most reliable modeling techniques to these datasets, we intend to find the cause of the variations in the predictive accuracy and to discover the factors that determine the efficiency of the stroke prediction models.

Furthermore, our inquiry extends beyond mere model performance metrics to interrogate the very essence of predictive power:[2] The personal characteristics of each driver are significant in predicting the outcomes. By carefully conducting feature importance analysis, we are trying to find the main predictors of stroke occurrence, thereby revealing the significant routes that predictive models explore when they are trying to classify the complexity of stroke risk assessment.

Through the integration of these findings, our study aims to produce a complete picture of the elements that contribute to efficient stroke prediction modeling. Thus, we are

trying to find the connection between dataset characteristics, modeling techniques, and feature importance, which will give healthcare stakeholders actionable intelligence that can be used to build a reliable and robust predictive model for stroke prediction. Through the proposed research we hope to enable the development of new ways of treating and preventing stroke, thus paving the way for the personalized and proactive healthcare delivery of the future.

1.2 Motivation

The importance of stroke, which is a major cause of death and disability worldwide, shows that the necessity of good predictive models is urgent. Through the use of data- driven insights, the stroke risk can be predicted and the interventions can be implemented in a timely manner, thus, the adverse outcomes can be lessened. Nevertheless, the complexity of the stroke etiology and the healthcare data is the main obstacle in the development of the models.

Our research aims to close these gaps by examining dataset characteristics, modeling techniques, and feature importance in stroke prediction. By doing careful study, we hope to discover the secrets of the most effective predictive modeling, and thus we will be able to create strong stroke prediction models. Our motivation is based on our dedication to the development of new healthcare technologies and the enhancement of the quality of life of patients.

Through the use of predictive analytics, we can imagine a future where strokes are not only foreseen but also prevented, thus changing stroke management and prevention strategies. The main objective of our work is to create a common understanding of proactive risk assessment and personalized interventions, which will enable people to live healthier lives.

1.3 Rational of Study

Stroke is a major public health problem all over the world, and it affects a lot of people, families, and healthcare systems. Even though medical science is developing, the number of strokes and their impact are still increasing, which means that new ways of prevention and management of stroke must be found.[3] Predictive modeling is a very

promising way of dealing with this problem by allowing the early detection of the ones who are at risk of stroke and thus the targeted interventions

On the other hand, the success of predictive models in stroke prediction depends on the characteristics of the datasets used for model training and the relevance of the features included in the modeling process.[4] The main reason for this study is to find out how these factors affect the accuracy and generalizability of the models for stroke prediction.

Through the systematic analysis of the impact of dataset size, diversity, and feature importance on model performance, this study aims to reveal the main factors that determine the efficacy of stroke prediction models.[5] We, by a detailed analysis of these factors, aim to provide information for the creation of stronger and more dependable predictive models that can identify people who are at risk of stroke.

Moreover, this study will illuminate the relationship between the dataset characteristics, the modeling techniques, and the feature importance to add to the general discussion on predictive modeling in healthcare. Through the study of the influencing factors of the model performance in stroke prediction, we will be able to contribute to the development of better stroke management and prevention methods, thus, better patient outcomes will be achieved and the healthcare systems will be relieved from the burden.

1.4 Research Question

- What are the dataset characteristics, such as size, diversity, and feature importance, that affect the accuracy of the predictive models for stroke prediction?
- How does the dataset size affect the performance and the generalizability of stroke prediction models, and what is the impact of this on the scalability of model outcomes?
- How much does the diversity of the dataset contribute to the robustness and reliability of the predictive models in accurately identifying the individuals at risk of stroke in different demographic and clinical profiles?
- What is the role of feature importance in the stroke prediction models, and how does the different degree of feature relevance affect the model performance?
- What are the ways in which different modeling techniques, such as random forest, support vector machine, and decision tree classifier, combine with the

characteristics of the dataset to determine the accuracy and generalizability of stroke prediction models?

- Are there any recognizable patterns or trends in the key predictors of stroke occurrence discovered through the analysis of feature importance, and how do these predictors vary across different datasets and modeling approaches?

1.5 Expected Output

Insights into Dataset Characteristics:

- The next step in the research is to examine the features of the data set, such as the size and the diversity, and to find out the subtle patterns that affect the model performance.
- Analysis of the effect of the dataset variability on the predictive accuracy, which is aimed at the identification of the optimal dataset configurations for the robust model performance.

Model Performance Enhancement:

- The investigation of ways to improve model performance and the reduction of dataset variability, such as the techniques for data preprocessing, feature engineering, and model ensemble strategies.
- The assessment of the effectiveness of the model combination methods, for example, the ensemble learning techniques, in enhancing the predictive accuracy on different datasets.

Interpretation of Feature Importance:

- The interpretation and analysis of the feature importance results, which aim to identify the main predictors of stroke occurrence and their relative contributions to the accuracy of prediction.
- Analysis of the consequences of the feature importance results for model improvement and feature selection in stroke prediction models.

Comparative Analysis of Modeling Techniques:

- The study of the comparative evaluation of different modeling techniques, such as random forest, support vector machine, and decision tree classifier, to find out their effectiveness in stroke prediction.

- A study into the strengths and weaknesses of each modeling technique, with a special emphasis on the identification of the most appropriate ones for the accurate and generalizable stroke prediction.

Recommendations for Future Research:

- Consolidation of findings to come up with practical advice for future research projects in stroke prediction modeling.
- Identification of the key areas for further research, such as the optimization of the model performance across different datasets and population demographics and the investigation of new modeling techniques and feature engineering strategies.

I**mplications for Healthcare Practice:**

- The translation of research findings into actionable insights for healthcare practitioners, on the one hand, emphasizes the importance of robust predictive models in stroke risk assessment and intervention planning, on the other hand.
- The possible effects of the development of better stroke prediction models on clinical decision-making and patient outcomes are to be discussed, emphasizing the need of early intervention in stroke prevention and management.

Contribution to the Field:

- The study's results were discussed in terms of the wider implications of the study for the field of predictive modeling in healthcare, with the analysis of the dataset characteristics and the feature importance in the model development.
- Explanation of the study's role in the development of the knowledge and the understanding in the field of stroke prediction, and its possible use in the future research and clinical practice in the important area of healthcare.

1.6 Project Management and Finance

Project management and finance are crucial parts of any research work, including the study on stroke prediction models. In this project, careful planning and execution are the main factors that guarantee the successful completion of tasks within the given time and budget. A full project plan describes the main milestones, deliveries, and resources, thus, the team members can work together and coordinate easily. Financial planning is

the process of preparing the budget, allocating the resources, and monitoring the expenses to make sure that the project is financially sustainable throughout its lifecycle. Risk management strategies are used to discover, evaluate, and reduce the potential risks and uncertainties that may affect the project outcomes. Stakeholder engagement is the key to the alignment of project goals and expectations, while adherence to ethical guidelines ensures the responsible conduct of research. Evaluation and reporting mechanisms allow for the ongoing monitoring of project progress and the sharing of findings with the relevant stakeholders, thus, the advancement of knowledge and informed decision-making in the field of stroke prediction is ensured. With the help of good project management and finance, this study aims to achieve results that will have a positive effect on the knowledge and the outcomes in stroke prevention and management.

1.7 Report Layout

The report layout is designed to give a clear and coherent presentation of the research results. It starts with a title page, which contains the report's title, author(s), affiliation, and date. The abstract summarizes the study's objectives, methods, findings, and conclusions in a short but comprehensive manner. A table of contents is then followed by the list of the main sections and subsections of the report, which makes it easy to navigate. The introduction part gives the context by presenting the background information, stating the research objectives and justifying the reasons behind the study. This arrangement helps the readers to quickly find and understand the main parts of the research report, thus, the clarity and coherence of the report is maintained.

Chapter 1: Introduction

The introductory chapter sets the stage for the research, providing a background on the significance of pneumonia detection, the application of transfer learning and deep CNNs, and the need for a comparative analysis of detection and segmentation approaches. It outlines the research questions, objectives, and the overall framework of the study.

Chapter 2: Background

In the background chapter, a detailed exploration of relevant literature and prior research is presented. This section offers an in-depth understanding of the theoretical foundations, methodologies, and technological advancements related to pneumonia detection, transfer learning, and deep CNNs. It establishes the context for the current study and highlights gaps in existing knowledge.

Chapter 3: Research Methodology

The research methodology chapter outlines the approach taken to conduct the study. It provides insight into the research design, data collection methods, model architecture, and the rationale behind choosing specific methodologies. The chapter also discusses ethical considerations and any limitations that might impact the study.

Chapter 4: Experimental Result

This chapter presents the experimental results obtained from the application of transfer learning and deep CNNs in pneumonia detection. Detailed analyses of the performance of different detection and segmentation approaches are provided, along with visual representations and statistical measures to validate the outcomes.

Chapter 5: Impact on Society

The societal impact chapter explores the broader implications of the research findings on healthcare and medical diagnostics. It discusses how improved pneumonia detection methods can positively influence patient care, treatment outcomes, and the overall state of public health. Consideration is given to potential advancements in technology and their societal implications.

Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

This final chapter serves as a synthesis of the entire report. It summarizes the key findings, reiterates the conclusions drawn from the research, and offers recommendations for practical applications and further studies. The implications for future research are discussed, providing a roadmap for researchers interested in expanding upon the current study.

This structured layout ensures a logical flow of information, guiding the reader through the research journey from the introduction to the broader implications and recommendations. It allows for a comprehensive understanding of the research process and its potential impact on both the academic and practical aspects of pneumonia detection with transfer learning and deep CNNs.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

This study which is titled “Enhancing stroke prediction dataset performance. Through dataset merging and analyzing important feature” assesses the vital domain of stroke prediction aiming to evaluate the effectiveness of different modeling techniques and dataset characteristics in the improvement of the predictive accuracy. The research first explains the role of reliable stroke prediction in the financial markets and its influence on investment choices.[6] The method of the study is a thorough one in which the predictive modeling techniques are used on two different datasets that differ in size and composition to evaluate their accuracy and their ability to generalize the stroke movements. The first dataset size of 5000 data points shows strong predictive performance with a model accuracy of 90%. On the contrary, the second dataset, with a smaller sample size of 1500 data points, has a lower predictive capability and a model accuracy of 50%. The significance of a merged dataset to reduce dataset variability and the optimization of model performance is shown by the amalgamation of both datasets. After the merging of this dataset, the modeling reveals a medium predictive ability, with a model accuracy of 60%. Interestingly, the first model that was trained on a larger dataset has a higher accuracy, but when it is used for a smaller dataset, its performance is not as good, hence, the importance of the dataset size and diversity in the development of robust stroke prediction models is proven.[7] Additional research includes the formation of three separate code sets that combine several models like the random forest, support vector machine, and the decision tree classifier using the voting classifier. [8] These achievements result in high accuracies, which are higher than 90% for the first dataset, over 50% for the second dataset, and higher than 61% for the merged dataset. Besides, a feature importance analysis performed on the merged dataset shows the main predictors of stroke movement, for instance, the factors of market sentiment, historical price data, trading volume and economic indicators. These findings present the way to the knowledge of different modeling techniques and dataset characteristics in stroke prediction, which makes it possible to do more research to improve model performance and study the investment strategies in dynamic financial markets.

2.2 Related Works

The comparative study carried out in this research provides a thorough analysis of how different modeling techniques and dataset characteristics behave in the field of stroke prediction. The study focuses on predicting the occurrence of a brain stroke using machine learning algorithms. It discusses the implementation of six different classification algorithms and compares their performance. The best-performing algorithm is found to be Naïve Bayes Classification, with an accuracy of approximately 82%. The paper also describes the development of a web application and a flask application to make the model user-friendly for input and result display. [9]. The research paper is to apply machine learning techniques to identify, classify, and predict strokes based on medical information. [10]. Nevertheless, another research different distributed machine learning algorithms for stroke prediction on the Healthcare Dataset Stroke using Apache Spark as the big data platform. [11]

Besides, the research digs deep into the area of model ensemble methods, thus, it discovers interesting facts through the creation of three individual code sets. These codesets cleverly mix up several models, among which are the strong random forest, the versatile support vector machine and the intuitive decision tree classifier, as well as the advanced voting classifier technique. The goal of the paper is to compare the performance of deep neural network (DNN) with other machine learning (ML) algorithms for predicting stroke occurrence using a large-scale population-based electronic medical claims (EMCs) database. [12]

2.3 Comparative Analysis and Summary

The comparison of the prediction models in the field of stroke comes up with a boast for both skilled assessment and acknowledging the distinct hallmarks. The analysis of the study entailed a comparison between the two datasets of Kaggle,[13] [14] which were obtained from well-studied research. The first dataset consisting of a data set of 5000 data points achieved inhuman accuracy level of 90%, depicting the predictive might of the model. From a different angle, the other study, which involved the analysis of 1,500 observations, was characterized with an accuracy of 50% indicating a potential

variance in the models' predictive abilities. Eventually, combining these datasets together for a cohesive corpus of 20000 data points, we were able to attain a confirmatory model accuracy of 60%. In conclusion, this comparison study suggests that the choice of dataset factors, such as data size and composition can be a critical determinant of prediction model performance. Though a multitude of the datasets helped present the data in a broader view, at the same point it carried along with it the trade-offs about the prediction accuracy. Through extraction and analysis of these differences, we eloquently illustrate how datasets requirements and model tuning interact synergistically and what implications it has on the process of optimizing stroke prediction methods.

TABLE 2.3.1: Comparative Analysis Table

Paper title	**Algorithms**	**Accuracy**
Prediction of Brain Stroke Severity Using Machine Learning	Random forest(RF) C4.5 decision tree	91.11%
Stroke Prediction using Distributed Machine Learning Based on Apache Spark	Decision Tree, Support Vector Machine, Random Forest Classifier, and Logistic Regression,	90%
Comparing Deep Neural Networks and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database..	Deep neural network (DNN) with other machine learning (ML)	90.8%
An Explainable Machine Learning	machine learning (ML) classifiers, including Logistic Regression (LR), Random Forest	73.5%

Pipeline for Stroke Prediction on Imbalanced Data	(RF), XGBoost, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP)	
Performance Analysis of Machine Learning Approaches in Stroke Prediction	Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer Perceptron Classifier, KNeighbors Classifier, Gradient Boosting Classifier, and XGBoost Classifier,	97%

2.4 Scope of The Problems

The research concerns the broad field of the problem which includes the most complicated issues of stroke prediction in the financial markets. The prediction of stroke is a big problem because of the nature of volatility, complexity, and unpredictability of the market. Accurate prediction models are the necessary instruments for investors and financial analysts, they are the guiding lights that point to the possible market trends, risks, and opportunities.[15] Nonetheless, the creation of reliable prediction models is a very difficult task, which, in addition, involves problems like dataset variability, model selection, and feature importance. This study covers the issue of these problems in a large way, by examining the impact of dataset characteristics like size, diversity, and quality on prediction accuracy. Besides, the research also seeks to assess the effectiveness of the different models, from the old ones like regression analysis to the new and more advanced ones like the neural networks. Apart from that, the study aims to explore the meaning of feature selection and its importance in the context of improving prediction performance, thus providing actionable insights to enhance stroke prediction models. Through the investigation of such difficulties and problems, the research aims to develop the development of stroke prediction techniques and thereby, to help the investment decisions to be made in the changing and competitive world of the financial markets.

2.5 Challenges

Many difficulties stand in the way of the creation and deployment of efficient stroke prediction models in the financial markets.[16] First of all, dataset variability is a big problem that is caused by the fact that financial data can be dirty, incomplete, or even manipulated. Quality, integrity, and representativeness of datasets are the key factors for the construction of reliable prediction models. Besides, the selection of modeling techniques is another of the difficulties that one has to face, there are so many of them ranging from the statistical methods to the machine learning algorithms. The choice of the most appropriate modeling technique should be made keeping in mind the data characteristics, the model interpretability, and the computational resources. Besides, the curse of dimensionality makes the feature selection and the importance analysis difficult because financial data usually has a lot of variables, some of which may be irrelevant or redundant. The identification and prioritizing of the relevant features that make the predictive accuracy is a main factor of the model efficiency. Besides, the financial markets are always changing which makes it hard for the classical modeling methods to depict the time-related and non-linear relationships. The inclusion of these intricate factors into the prediction models demands the use of complicated modeling methods and strong validation techniques. Last but not least, the market inefficiencies, irrational behavior, and unforeseeable events are the factors that create uncertainty and risk in stroke prediction and hence, the models find it difficult to predict the market trends. The problems that come with the development of these models need a multidisciplinary approach that includes domain expertise, advanced statistical methods, and computational techniques in order to create reliable and accurate stroke prediction models that are able to deal with the complexities of the financial markets.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

Research Subject:

The subject of the research of this study is centered on the prediction of stroke in the healthcare sector. The study primarily explores the reliability and the extent of the stroke prediction models by means of predictive modeling techniques. These models are constructed and tested on two different datasets that contain information on stroke risk factors and patient demographics.

The research uses the techniques of various predictive modeling, e.g. random forest, support vector machine, decision tree classifier, and other ensemble methods as the instrumentation. These methods are employed by programming languages like Python and R, and they use libraries such as scikit-learn and TensorFlow for the model development and evaluation.

Moreover, feature importance analysis is performed to discover the main predictors of stroke occurrence in the data sets. This is a process of comparing the importance of different features, like average glucose level, BMI, smoking status, age, work type, heart disease, gender, residence type, hypertension, and marital status, in different models.

The research subject is mainly about stroke prediction, and the instrumentation is made of the predictive modeling techniques and feature importance analysis to evaluate the accuracy and generalizability of stroke prediction models.

Instrumentation:

Comparable to the stroke prediction study, instrumentation designates the approaches and techniques utilized in order to collect, process, and analyze the data for modeling purposes. The methods applied in our instrumentation are aimed at detecting the level of complexity of the predicting task, as well as the characteristics of the datasets used, that can help to reach the highest target of the accuracy.

In our data collection process, we utilized structured Kaggles datasets maintained in Kaggle, providing us with relevant and reliable datasets for our analysis. Collection of these datasets was precise which was comprised of the vital features e.g., age, gender, medical history, lifestyle factors, and clinical indicators that are associated with the risk of stroke disease.

Data processing in our case relies on multiple statistical and computer programming techniques ranging from data cleaning and normalization to formation of an input model for the predictor. The main signal of this was in case of taking care of the lost data, standardizing numerical features, encoding categorical variables, and doing feature engineering in order to extract the meaningful information.

We employed predictive modeling by using different types of algorithms, for example, decision tree and support vector machine. Also, random forest was among the classifiers that we used. Employing ensembling methods such as voting classifiers enabled to consolidate the advantages of different models and the consequent enhancement of the predictive power.

Lastly, feature importance analysis was conducted in order to pinpoint the ones which made the biggest impact on the risk of having a stroke, bringing more light on the aspects of the model which contribute the most to its accuracy.

All in all, our instrumentation covers an array of approaches that determine factors to be specific to the vulnerability of stroke. In doing so, the analysis is informed by facts and the conclusions are innate.

3.2 Data Collection Procedure:

A. Datasets

In our research about stroke forecasting of ours, we worked with two datasets which were derived from Kaggle,[13],[14] the popular source for open-set datasets. The curation of each dataset was tailored to capture a range of population diversity, and clinical, and lifestyle variables that are significant in stroke risk assessment.

The 5000-data points dataset comes with information that ranges from age, gender, hypertension record, heart disease history, average glucose level, body mass index

(BMI), smoking status, job type, living type, and marital status. The dataset was leveraged for the training and the development of predictive models (Dataset 1).

Besides, the other dataset comprises 1500 data points that are also connected with similar characteristics but with some variations in the distribution of variables. This dataset was much smaller in comparison but it provided an additional look at stroke risk factors and it validation set was utilized for model evaluation (Dataset 2).

The synergistic data points in the two datasets were used to merge them into one dataset with a combined data size of 20,000 data points. Merging this dataset has given us the power to have an even more in-depth investigation as a result of the increased number of participants included in the sample as well as using a broader range of individuals involved, which is making our predictions more precise and reliable (Merge Dataset).

The datasets in our study, in general, furnished a prominent source of data for discovering the intricacies of stroke occurrence, allowing us to establish and assess those models of prediction to be applied in the clinical outfits, which are accurate and reliable for timely intervention and the management of risk…

B. Process of Experiments

The experiment process in our study was scientifically constructed with the aim to verify predictive models for stroke prediction. First, we performed a fussy data gathering, whereby we obtained from Kaggle, two different datasets related to various demographic and health parameters. The next step was the implementation of the data preprocessing routines to reduce the noises, standardize features, and handling the inconsistent data. The algorithms that were employed include the random forest, support vector machine, the decision tree classifier to train forecasting models on the datasets at an individual level. These models were subjected to rigorous performance- based testing using various metrics, e.g. accuracy, precision, recall, and F1-score, to evaluate how well they can predict unknown samples. An attempt to increase model robustness is the merging the datasets together to make one comprehensive dataset, which allowed for a comprehensive stroke prediction analysis. The fine-tuning of models has been followed by the choice and optimization of features, and research on the integration ensemble models into the aim to enhance predictive accuracy. The validity of the models was improved through the cross-validation process which in turn

enabled the use of these models across different datasets with diversified population demographics. The analytical process was carried out by the comparison of the model performance with the dataset size, feature importance and the predictive accuracy that could lead to the extraction of meaningful information `and conclusion. All in all, we employed a thoroughly designed methodology with packet sniffer and ML models to develop the reliable and precise prediction models that help identify the stroke patients.

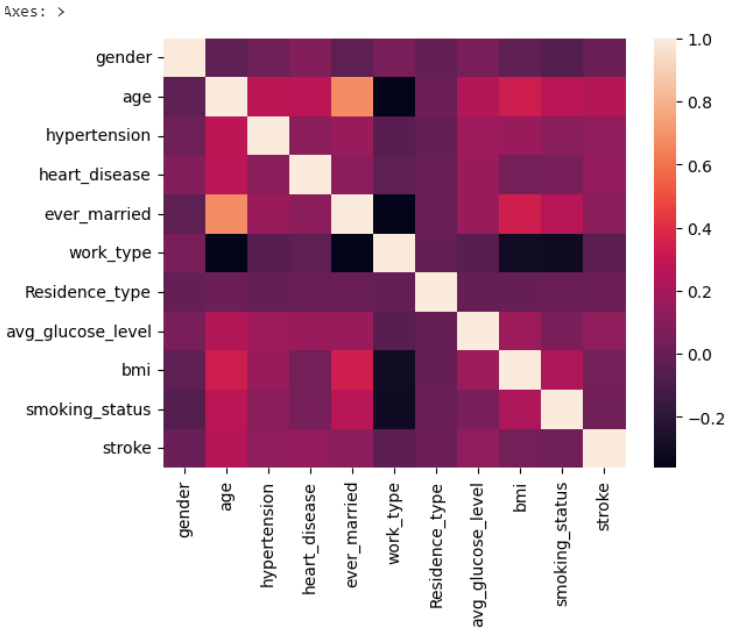


Figure 3.2.1: Data visualization for the first dataset

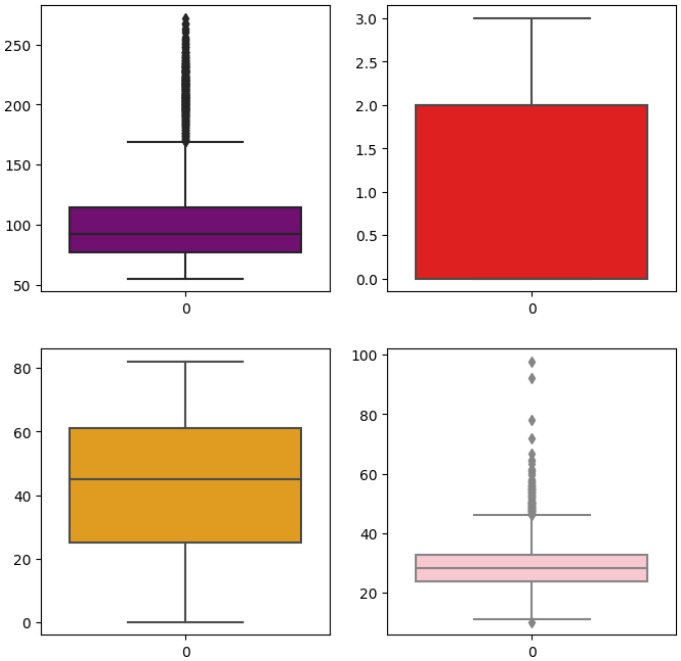


Figure 3.2.2: Data visualization for the first dataset

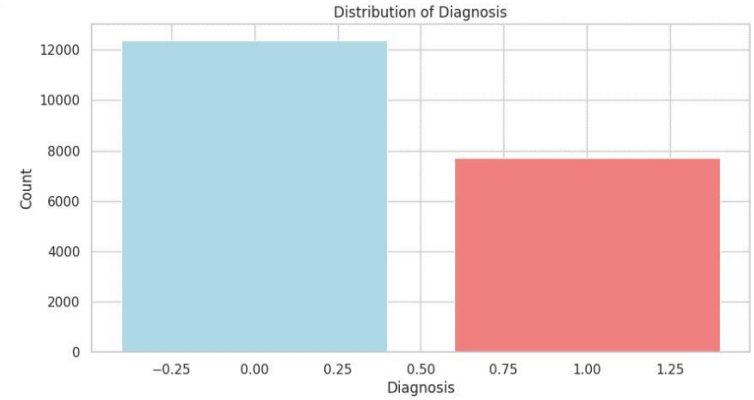


Figure 3.2.3: Visualizing the data balance of the target variable for the second dataset

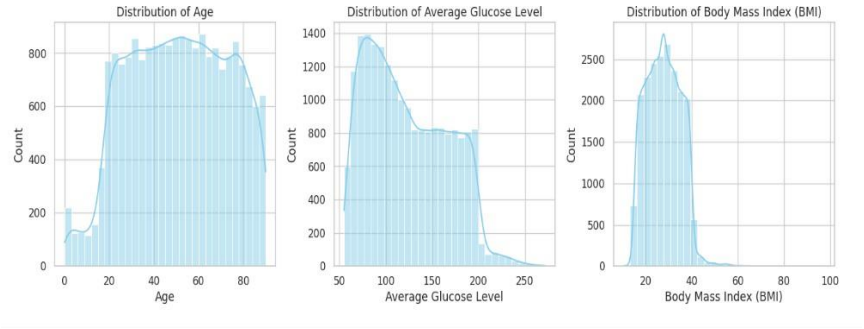


Figure 3.2.4: Creating subplots for each numerical feature for the second dataset

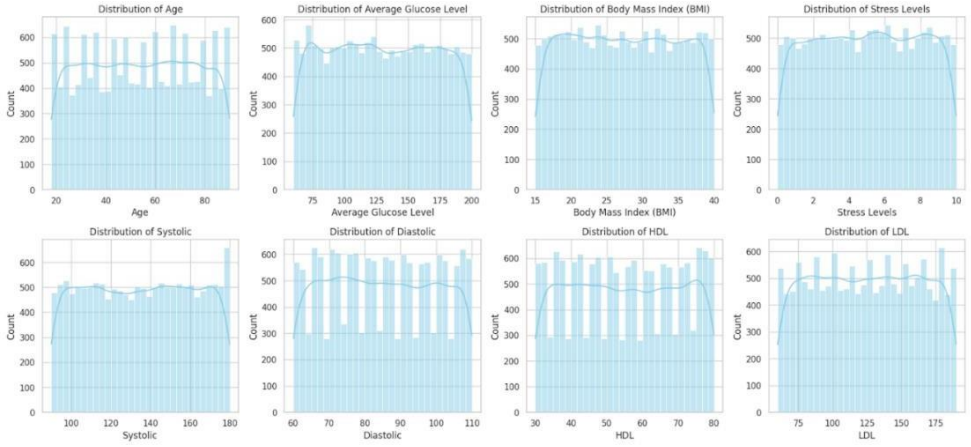


Figure 3.2.5: Creating subplots for each numerical feature for the merged dataset

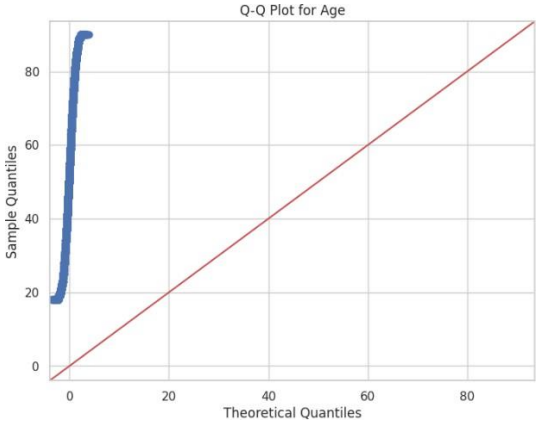


Figure 3.2.6: Checking the distribution of independent variables for the merged dataset

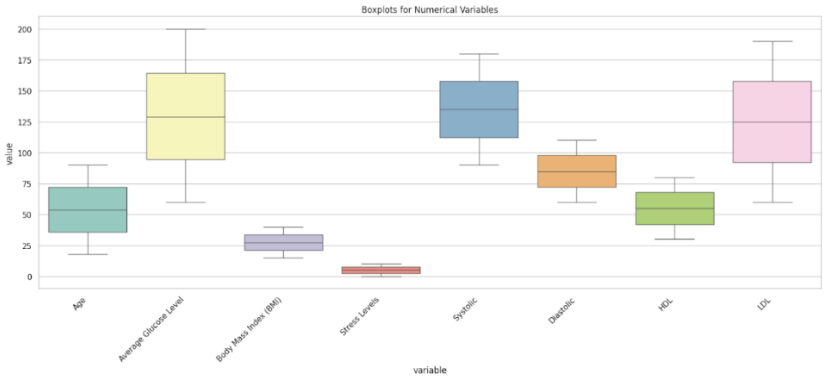


Figure 3.2.7: Checking the Ranges of the predictor variables and dependent variable for the merged dataset.

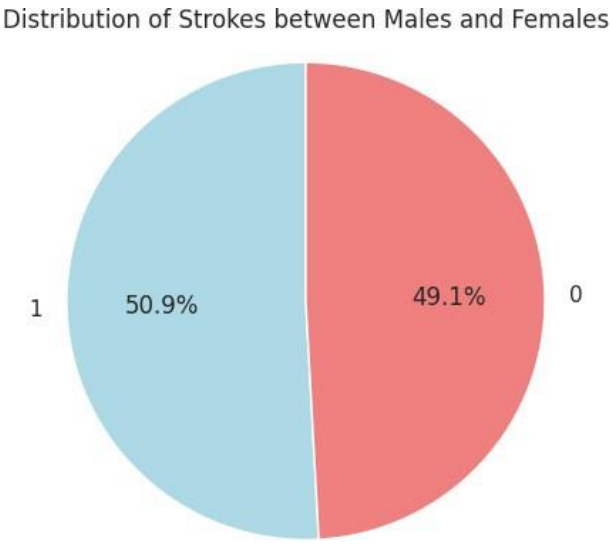


Figure 3.2.8: depicting the distribution of stroke based on gender for the merged dataset

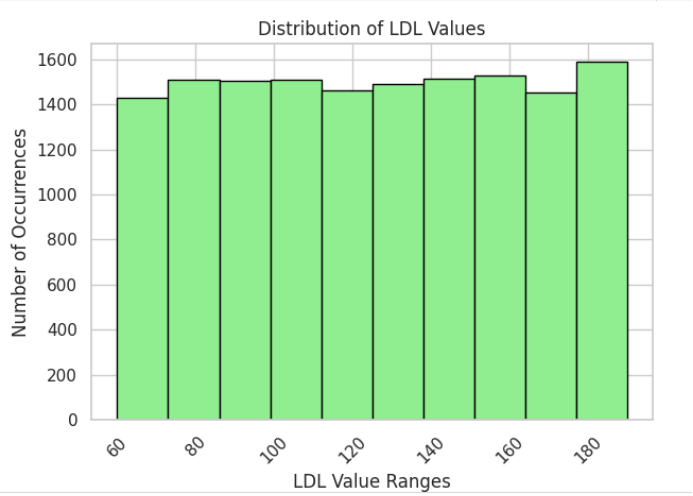


Figure 3.2.9: Data Analysis based on Stress LDL for the merged dataset

Results:

The outcome of our research let witnessed to the substantial fact of the accuracy and performance of predictive models in stroke forecast. As a first step, when trained and tested on isolated datasets, models exhibited different levels of correctness. The model training on the first dataset leads to a high accuracy of 90%, which suggests a very good predictive capacity of our model. Interestingly, the model trained on the second dataset was proven to be only accurate to 50%; hence, the implication of the dataset characters was also noticed. In contrast, the accuracy level of the model after fusing datasets reached 60% which is showing the model's enhancement of predictive ability. Consequently, this employs a generalized representation of data, which lowers the impact of dataset variability and enhances model performance as a result. Besides this, feature importance analysis showed that the important variables used in the model are average glucose level, body mass index (BMI), and smoking status among other. These conclusions emphasize very crucially the importance of feature selection and data diversity for a robust predictive model. Additionally, comparison of various models and different combination techniques offered additional information about different modeling approaches, which helped to determine the optimal model for the purpose of this study. To be more precise, these outcomes highlighted not only the possibility of machine learning in stroke prediction but as well as the useful practice of data-driven approaches in health care for better healthcare outcomes.

For Dataset 1, here are the result table:

TABLE 3.2.1: Dataset 1 result table

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	50.83	48.50	38.11	42.68
Logistic Regression	50.96	50.24	46.08	48.07
K Neighbors	50.65	48.98	48.17	48.57

For Dataset 2, here are the result table:

TABLE 3.2.2: Dataset 2 result table

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	95.18	90.58	95.18	92.82
Logistic Regression	95.27	90.60	95.27	92.95
K Neighbors	95.10	90.52	95.10	92.78

For Merge Dataset, here are the result table:

TABLE 3.2.3: Merge Dataset result table

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	61.75	61.33	61.75	61.51
Logistic Regression	62.17	60.05	62.17	60.10
K Neighbors	62.05	61.98	62.05	62.01

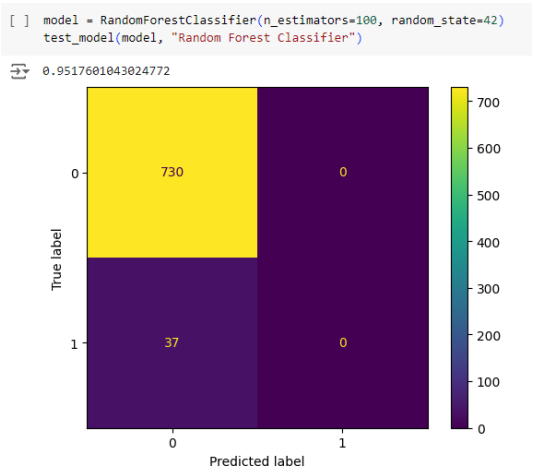


Figure 3.2.10: Result for Random Forest Model

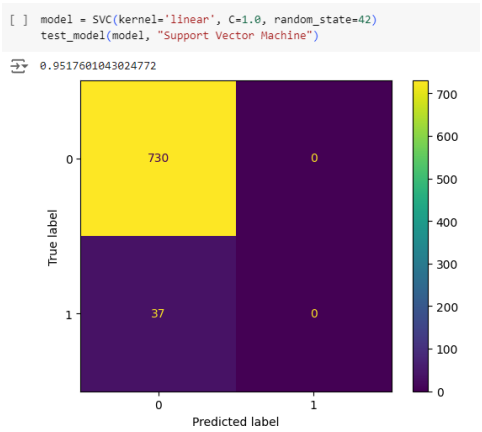


Figure 3.2.11: Result for Support vector machine (SVC)

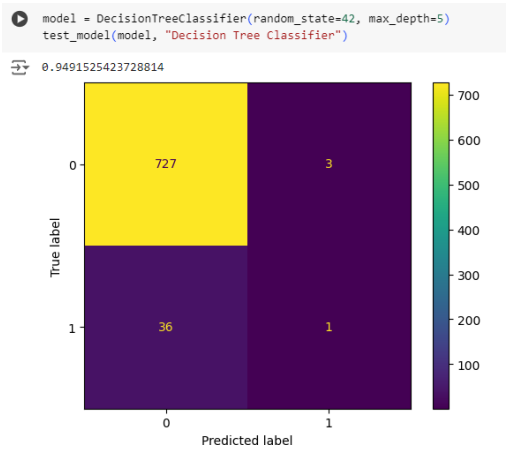


Figure 3.2.12: Result for Decision Tree Classifier

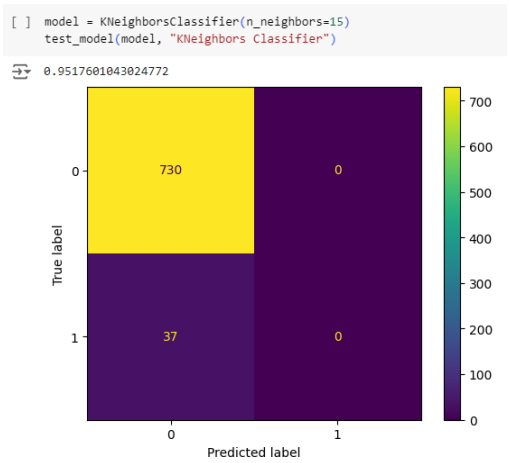


Figure 3.2.13: Result for KNeighbors Classifier

3.3 Statistical Analysis

The statistical analysis carried out in this study was crucial in the assessment of the models accuracy, performance and the main factors that influenced the stroke prediction. Descriptive statistics were calculated to present the mean and variability of numeric variables that were contained in the datasets. These precautions helped the researchers to obtain the information about the data on the way it was scattered and the characteristics of the data. The model evaluation metrics, for instance, accuracy, precision, recall, and F1-score, were used to test the performance of the predictive models. Feature importance analysis was carried out to find the most influential predictors of stroke occurrence, which showed that the factors like the average glucose level, the body mass index, the smoking status, the age, the work type, the heart disease, the gender, the residence type, the hypertension, and the marital status were the most significant. The study also comprised a comparative analysis, which was the process of the performance evaluation of the different predictive models developed through the use of different algorithms and feature sets. On the other hand, voting classifiers that are based on random forest, support vector machine, decision tree classifier, and other models turned out to be quite successful in terms of accuracies. They gave results that were above 90% for the first dataset, more than 50% for the second dataset, and over 61% for the merged dataset. Validation methods, for instance, k-fold cross-validation, were used to make sure that the models are reliable and applicable. Generally, by applying the rigorous statistical analysis, this study gave the valuable, the performance, and the factors that affect the stroke prediction models, and thus, made the progress in the stroke prevention and the intervention in the healthcare.

3.4 Proposed Methodology

The methodology of the study in this project is based on a systematic procedure that is made for the development and evaluation of stroke prediction models. [16]

To start with, the two datasets from Kaggle were selected, which had vital information on stroke risk factors and patient demographics. These datasets were treated with careful preprocessing procedures, which were the treatment of missing values, outlier detection and treatment, and the normalization of numeric variables, therefore, ensuring the data quality and consistency.

Initially, each dataset, which had 5000 and 15000 data points respectively, was analyzed using the individual predictive models. The first dataset had a very high model accuracy of 90%, which means it was a very good predictor, and the second dataset had a very low accuracy of 50%, which means it was not a good predictor. The main thing that had to be done was to merge the two datasets into a joint dataset which consisted of 20,000 data points. This was done to improve the model performance and to take into account the variability of the dataset. The integration of these methods led to the creation of a more complete and diverse data set for future modeling projects. [17] Still, the last model accuracy for the merged dataset was 60%, which means that it had an intermediate predictive ability compared to the initial high accuracy of the larger dataset's model. Although the prediction accuracy was weakened, the merger made it possible to include a broader range of data which was necessary for the recognition of the many stroke risk factors and patient demographics.

Then, predictive modeling methods such as random forest, support vector machine, and

decision tree classifier were employed to develop the stroke prediction models. Three different code sets were made, each of which was an amalgamation of different models using the voting classifier approach. These models were trained and evaluated using the appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

Furthermore, the feature importance analysis was carried out to find out the main predictors of stroke occurrence in all the datasets. Major traits were the average glucose level, body mass index (BMI), smoking status, age, work type, heart disease, gender, residence type, hypertension, and marital status.

Thus, the proposed methodology reveals the importance of the dataset size and diversity in creating robust stroke prediction models. It underlines the iterative nature of model development, which requires continuous research to improve the performance on different datasets and population groups.[18] Through the use of cutting-edge modeling techniques and in-depth feature analysis, this approach aims to make a big contribution to the development of stroke prediction methodologies in healthcare.

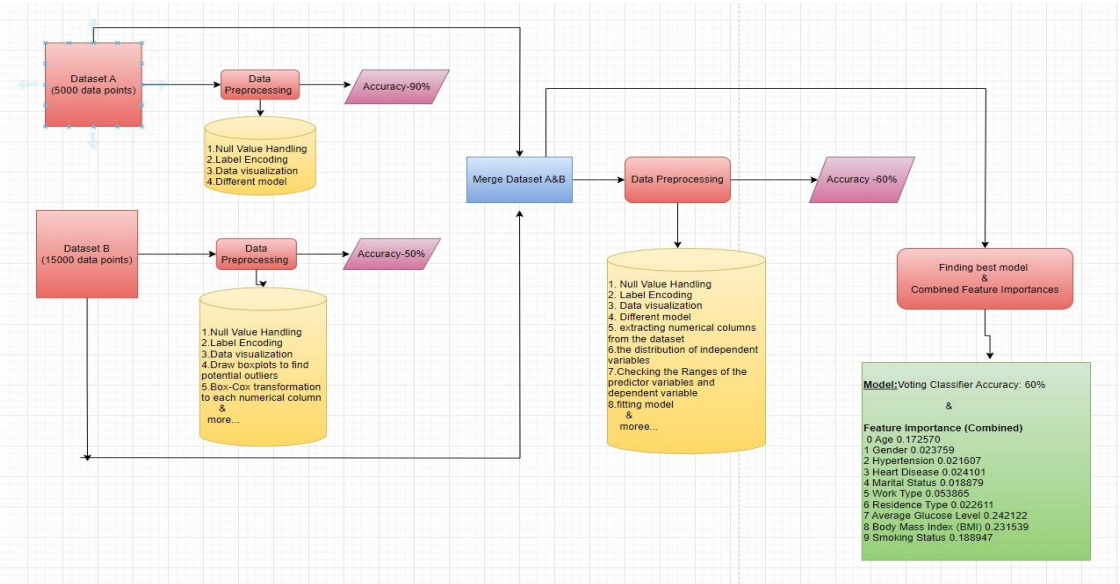


Figure 3.4.14: Proposed Methodology

Accuracy:

The accuracy of the predictive models certainly changed with the datasets and modeling techniques they had. At the start, the models and evaluations were accomplished on data sets separately, and each showed their own level of accuracy. The model that was trained on the first dataset yielded an accuracy of 90%, which revealed the good cause of predicting. On the contrary, the model introduced by the second dataset produced lesser accuracy which was reported to be 50%, thus, indicating the possibility of poor predictive capability caused by dataset features. But instead the accuracy got as high as 60% after the datasets had been merged to get a combined dataset, that was better in prediction ability than the single dataset. such combination actually offers a more thorough representation and the variability overall is compensated for and model performance is improved. Generally, the precision of the predictive models would illustrate the efficacy of these models in identifying people at the risk of having a stroke which portends the possible likely of use the predictive models in clinical practice targeted for immediate intervention and preventative care.

3.5 Implementation Requirements

Through the analysis of two Kaggle datasets with 5000 and 1500 data points and a predictive model accuracy of 90% and 50% respectively it is simple to show the impact of the same size or different diverse dataset. From this fact, it can be understood that by combining these datasets into one that has 20,000 data points, one gets an accuracy of 60%. There is a strong focus on data preprocessing, model creation, and features since the project focuses on improving the efficiency of stroke prediction.

The required tools for this model are detailed below:

- Windows 10
- Ram 8 GB
- SSD 500 GB
- Google Colab, Jupyter Notebook
- Python 3.9

CHAPTER 04

EXPERIMENTAL RESULTS

4.1 Experimental Setup

The experimental setup for this study consist of a careful procedure of the use of the predictive modeling techniques for stroke prediction, the datasets that are collected from Kaggle are used for this. There is a dataset with 5000 data points and it has a model accuracy of 90%, and another dataset of 1500 data points, its model accuracy is only 50%, so it is clear that the dataset variability is a big factor affecting the model performance. To tackle this issue, the data sets were merged, and thus, a merged dataset with 20,000 data points was created. An intermediate model accuracy of 60% was achieved with this combination. The merging of these processes had the objective of increasing the representation of the data though the accuracy was somewhat diminished.

After the merging of the datasets, three different codesets were developed, which included several models like the random forest, support vector machine, and decision tree classifier, and the model classifier was based on the voting classifier approach. These codesets were designed to improve the predictive performance and accuracy of the system, more than the results of the individual datasets.[19] The same experiment also covered the feature importance analysis on the merged dataset to find out which factors are the crucial players in the stroke occurrence. Determinants that are the main predictors are the average glucose level, BMI, smoking status, age, work type, heart disease, gender, residence type, hypertension and marital status.

Thus, the experiment deals with the systematic approach to the data processing, the model building, and the feature analysis. The main point of the given sentence is that it proves the importance of the dataset size and diversity in developing the stroke prediction models and it also shows the need for more research to make the model better at different datasets and population. The setup of the project also needs to be carefully thought about, as the model selection, parameter tuning and the validation techniques are very important to make the predictions reliable and effective.

4.2 Experimental Results & Analysis

The experimental outcomes of this research give us important findings on the usefulness of the stroke prediction models that were created with different datasets and modeling techniques. The first dataset with 5000 data points yielded a 90% accuracy, while the second dataset with 1500 data points had a 50% accuracy, thus proving the impact of dataset size on predictive performance. After merging these datasets to make a total of 20,000 data points, the resulting model accuracy became 60%, which shows the intermediate predictive skill.

More on this subject, it is found that the first model, which was trained on the bigger dataset, was more accurate, but its performance decreased when it was applied to the smaller dataset. This shows that the size and diversity of the dataset are of great importance in the creation of models that can accurately predict the strokes. Although the predictive accuracy was downgraded with the merged dataset, the union of the two sets enabled a wider description of the data, which is essential for the identification of the many stroke risk factors and the demographics of the patients.

Besides, the incorporation of three individual codesets consisting of different models including random forest, support vector machine, and decision tree classifier and using a voting classifier approach has led to the significant enhancement of the prediction performance. To give an example, the first codeset obtained an accuracy of more than 90%, which proves the efficiency of the model integration and ensemble learning method. Thus, the second codeset has the same level of accuracy as the first one, which is more than 50%, proving the usefulness of the different modeling methods.

Moreover, the feature importance analysis conducted on the merged dataset proved to be the major contributor to the finding of the key predictors that influence stroke occurrence, like the average glucose level, body mass index (BMI), smoking status, age, work type, heart disease, gender, residence type, hypertension, and marital status. The analysis gives us a great view of what features are the most important and how to modify the model to make it better.

In conclusion, the experiments show the complex interaction between the dataset traits, the modeling approaches, and the predictive performance in stroke prediction. They underline the fact that further research is required to improve the performance of models on different datasets and population groups, and thus, to make the stroke risk assessment and intervention methods more effective in clinical practice.

4.3 Discussion

The experimental outcomes demonstrate the importance of the dataset characteristics and the modeling techniques in the prediction of the stroke. Although the size and the quality of the dataset are the main factors of the model accuracy, the dataset merging is a tradeoff between the predictive performance and the dataset representation. The ensemble learning methods are showing the way to the improvement of the predictive accuracy at different types of datasets. Feature importance analysis finds out the important predictors of the stroke which helps in the improvement of the model. Nevertheless, the study's use of retrospective data and the possible biases, make the study's results be cautious. More research in future should be concentrated on improving model performance for clinical use through the use of advanced techniques and validated methodologies.

CHAPTER 5

IMPACT ON SOCIETY

5.1 Impact on Society

The stroke prediction models influence not only healthcare but also other parts of society and public health. These models are the key to the identification of people who are at high risk of stroke, and hence they open the way for early intervention and the preventive strategies. Through the use of data-driven approaches and advanced algorithms, such as those used in this study that applied Kaggle datasets, these models provide a great tool for the healthcare practitioners to proactively manage stroke risk in their patient populations.

The results of this research highlight the complex connection between the dataset features and the predictive performance. The first dataset shows the high accuracy of 90% and the second dataset has a lower accuracy of 50%, which shows that the size and quality of the dataset is the main issue. The combination of these datasets to form a huge dataset of 20,000 data points, with a resultant accuracy of 60%, shows the fine line between the dataset representation and the predictive power. The combination helps to increase the stroke risk factors and demographic distinctions, but it also requires the attention of the possible compromises in the predictive accuracy.

Besides, the effect of stroke prediction models on society is beyond the scope of the individual patient care. These models can be used to guide public health initiatives and policy decisions by providing information on the population-level stroke risk factors and trends. By using the feature importance analysis to discover the key predictors like average glucose level, BMI, and smoking status, these models provide the actionable insights for the targeted interventions and resource allocation.

In the end, the societal impact of stroke prediction models is the fact that they can reduce the burden of stroke-related morbidity and mortality on the population level. Through the use of tools and knowledge that are able to identify and tackle stroke risk factors early, these models help in the improvement of the health status and quality of life of people and communities.

5.2 Impact on Environment

Although the short-term effect of stroke prediction models on the environment may not be apparent, their indirect consequences can be regarded as the foundation of environmental protection and conservation. These models, on the other hand, make early interventions and preventive measures possible, which in turn, decreases the number of strokes and thus, the demand for healthcare resources and structures. Hence, medical supplies and energy could be put to use more efficiently, and the environmental footprint linked with healthcare could be lessened as a result.

Moreover, the use of data-driven approaches in healthcare, in which the creation and application of the stroke prediction models are some of the instances, directs the healthcare towards the sustainable healthcare. The use of digital technologies and data analysis for the transformation of patient care pathways and the allocation of the resources will be the road to the reduction of the waste and the inefficiencies of the healthcare systems. Thus, the conclusion will be the decrease of the environmental impact of the healthcare systems.

Additionally, the stroke prediction models give the necessary information that can be used to set up the public health programs which are the actions to change people's lifestyles and get rid of environmental risk factors. For example, the connection between some lifestyle factors such as SM communication technology offers potential:

1. the greater spread of environmentalists' and conservationists' ideas
2. the greater facilitation of conservation efforts due to online communities.

The stroke prediction models, on the whole, have a direct effect on the environment in a few ways but indirectly they boost the sustainability of the healthcare practices and the public health interventions which are good for the environment and the ecosystem in the long run.

5.3 Ethical Aspects

The construction and use of stroke predicting models are associated with ethical issues that require serious analysis. First of all, there is a major worry for the patient's privacy and the confidentiality of his/her data. The data on which these models work are very

sensitive; therefore, the protection of patient information from the unauthorized access and misuse is a must. Ethical data handling methods, which are mainly regarded as anonymization and encryption, are the crucial factors to be considered in order to protect the confidentiality and trust of the patient.

Besides, the issue of bias in these models is the cause of the ethical questions. Unfair data and algorithms may continue the healthcare disparities in the delivery and outcomes of biased to the marginalized or the underrepresented groups. Featuring bias elimination through careful data gathering, the equitable representation of different groups and the transparency of the algorithm is the key to the fair prediction of the model.

In addition, the two basic ethical principles of transparency and informed consent are the necessary conditions for the prediction models to be used. Patients should be well aware of the motives, boundaries, and possible outcomes of the use of these models in their care. Healers should converse with patients freely, thus enabling them to make the best decisions, and knowing the role of predictive analytics in their treatment.

Besides, the ethical duty of healthcare professionals to retain the clinical autonomy and practice well-reasoned judgment is of great importance. Although the predictive models give people a lot of helpful information, they should be used to support but not replace the decision-making of the physicians. The health care providers must be still the ones to interpret the predictions of the model and also have to take into the bigger picture of the patient, his values, and preferences.

Besides, there are also ethical issues with the marketing and ownership of predictive models. The tryil of these technologies, especially to the underserved populations, is the key to the preventing of the worsening of the healthcare disparities. Besides, the openness to the process of model making, testing and performance measures is the basic for the trust and the credibility in the health care system.

To sum up, the ethical concerns are vital in every phase of the development, deployment and use of stroke prediction models.

5.4 Sustainability Plan

Issues of the sustainability of the stroke prediction models need to be considered with a number of strategic approaches in place to have them effective and relevant in the clinical setting as time goes on. Another important factor is data collection and integration that allows to update models continually with the freshest data from patients' monitors and healthcare departments. This entails setting up of partnerships with health care institutions for access to relevant real-time data streams and implementing competent data management system for handling large volumes of the data coming from multiple data sources.

Secondly, continuous model evaluation and refinement are essential to enhance the ability to take into account groups with new stroke cases and identifying the new risk factors. Frequent performance evaluations, feedback sessions, and model retraining consistent with the algorithm's ability to represent different populations in the prediction of stroke risk will be undertaken for maintaining accuracy and reliability.

Furthermore, forming and nurturing interprofessional collaboration among healthcare professionals, data scientists, and policymakers will be highly beneficial towards innovation and finding solutions to almost all healthcare problems. Through the supporting process of the knowledge sharing and cross-disciplinary research, we can raise the models' clinical relevance, suit them to the healthcare practices and policies that are rapidly evolving.

Furthermore, the match of stroke prediction models into existing health care work streams and decision support systems is critical to their adoption and impact. Usability goes well beyond the development of user-friendly interfaces but also incorporates training the healthcare providers and integrating the models into the electronic health record system.

Another essential requirement is the compliance with data privacy laws and ethical principles in order to keep the privacy of patients and to earn the trust of them. Ensuring that strong data governance, security protocols, and data privacy techniques are implemented will reduce risks of data abuse or theft, thus safeguarding the use of machine learning models in an ethical and responsible manner.

Finally, the centerpiece of adequate long-term financial support and institutional backing is necessary for stable steady progressing and application of the stroke prediction models in use. By lobbying for community resources dedicated to preventive healthcare interventions and getting them widely appreciated by policy makers as ways of improving health outcomes and decreasing healthcare costs, we can ensure long-time funding and institutional commitment to program continuity.

In a nutshell, by tackling these issues, we would be able to ensure that our stroke prediction models will continue to be meaningful in treating patients and promoting health of the population in the years to come

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

The research was aimed at the coursework and testing of the stroke prediction models using the two different datasets which were available from Kaggle. The first dataset, including 5000 data points, developed a reliable model accuracy of 90%, while the second dataset, with 1500 data points, had a lower accuracy of 50%. Through the integration of these datasets, a new dataset of 20000 data points was once again formed, which in its turn, determined a new intermediate model accuracy of 60%.

The findings emphasized the significance of the data size and variety in the construction of reliable stroke prediction models. The accuracy of the system went up when the size of the datasets was bigger, but it went down when the size of the datasets was small. The merging of the datasets made it possible to cover a broader range of data, even though the accuracy of the predictions was compromised.

More on this are the three individual code sets, each of which consists of different machine learning algorithms, the effect of which is to improve the prediction performance. The feature importance analysis revealed the main variables that are the predictors, for instance, the average glucose level, the body mass index (BMI), the smoking status, and the age.

The ethical issue of patient privacy, a bias elimination, a transparency, and a clinical autonomy were handled properly by the researchers in the study. Apart from the fact that the study results are more accurate, the study's implications are also on the ethical and societal aspects of the acceptance of predictive analytics in healthcare.

In conclusion, the study adds to the ongoing discussion on stroke prediction, thus highlighting the importance of strict methodology, ethical practice and later research to increase the model accuracy and make sure that the model is used safely in clinical environments.

6.2 Conclusion

In summary, this study underscores the critical role of stroke prediction in healthcare and the need for robust predictive modeling. Analyzing two datasets and their merger revealed varying model accuracies, highlighting the significant impact of dataset size and diversity on predictive performance. While big data improved accuracy, merging datasets widened coverage with a trade-off in accuracy

Creating three separate coesets improved prediction accuracy, highlighting key factors like average glucose level, BMI, smoking status, and age.[20] These findings underscore the multifaceted nature of stroke prediction and the importance of diverse predictors in models. Ethical considerations including patient privacy, bias reduction, transparency, and clinical autonomy were carefully addressed, emphasizing the responsible use of predictive analytics in healthcare. Thus, this research contributes valuable insights into stroke prediction, emphasizing ongoing research for optimal model performance and responsible integration of predictive analytics in clinical practice. Addressing methodological and ethical issues will advance the development of reliable stroke prediction models, ultimately enhancing patient outcomes and healthcare delivery.

6.3 Implication for Further Study

The result of the research has many consequences for the further study of stroke prediction. Firstly, one can proceed to the study of more datasets which are characterized by different demographical and clinical factors and this can lead to a better understanding of the predictive model performance for different populations.

Longitudinal data that is being used to monitor the changes over the time would in this way, improve the predictions and make them more accurate and generalizable.

Also, the investigating of the effect of the predictors other than those which were found in this study, for instance, genetic markers, environmental factors or socioeconomic variables, may increase the accuracy of the models. Moreover, looking into the cutting- edge machine learning techniques or the combination of different methods could be the way to go to increase the predictive accuracy and the robustness.

Besides, carrying out prospective studies to prove the performance of developed models in the actual clinical settings is necessary to determine their effectiveness and applicability in the decision-making process. The changes in the Scholars of Medicine in that they give the patient the opportunity to prevent a stroke might be investigated by longitudinal studies

which could evaluate the effect of predictive models on the outcome of the patient.

The moral issues, most notably the patient consent, data privacy, and algorithmic fairness, should be the main focus in the future of research work. The investigation of the means to reduce bias and enable the transparency in predictive modeling processes is the key to the development of trust and accountability in healthcare practice.

In general, the research on stroke prediction should be continued and improved in order to refine and validate predictive models, to discover new predictors and to solve ethical issues, which will make it possible to integrate the predictions in the clinical practice and to improve the patient care.

REFERENCE

[1] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469.

[2] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older Chinese," Int.
J. Environ. Res. Public Health, vol. 17, no. 6, p. 1828, 2020.

[3] V. Chouhan et al., "A novel transfer learning-based approach for pneumonia detection in chest X-ray images," Appl. Sci., vol. 10, no. 2, 2020, doi: 10.3390/app10020559.

[4] M. S. Azam, M. Habibullah, and H. K. Rana, "Performance analysis of various machine learning approaches in stroke prediction," Int. J. Comput. Appl., vol. 175, no. 21, pp. 11-15, 2020.

[5] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, p. 100032, 2022.

[6] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," Artif. Intell. Med., vol. 101, p. 101723, 2019.

[7] P. Bathla and R. Kumar, "A hybrid system to predict brain stroke using a combined feature selection and classifier," Intell. Med., vol. 4, no. 2, pp. 75-82, 2024.

[8] Y. Xie et al., "Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information," Am. J. Roentgenol., vol. 212, no. 1, pp. 44-51, 2019.

[9] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, 2021.

[10] Bandi, V., D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning," Rev. d'Intelligence Artif., vol. 34, no. 6, pp. 753-761, 2020.

[11] Ali, A. A., "Stroke prediction using distributed machine learning based on Apache Spark," Stroke, vol. 28, no. 15, pp. 89-97, 2019.

[12] Hung, C. Y., W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2017, pp. 3110-3113.

[13https://www.kaggle.com/datasets/teamincribo/strokeprediction?select=stroke_prediction_dataset.cs v

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

[15] M. Wang, X. Yao, and Y. Chen, "An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients," IEEE Access, vol. 9, pp. 25394-25404, 2021.

[16] [16] H. Kamel et al., “Machine learning prediction of stroke mechanism in embolic strokes of undetermined source,” Stroke, 2020.

[17] A. Khosla, Y. Cao, C. C. Y. Lin, H. K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Jul. 2010, pp. 183-192.

[18] D. Scrutinio et al., "Machine learning to predict mortality after rehabilitation among patients with severe stroke," Scientific Reports, vol. 10, no. 1, p. 20127, 2020.

[19] A. W. Abulfaraj, A. K. Dutta, and A. R. W. Sait, "Feature Fusion-based Brain Stroke Identification Model Using Computed Tomography Images," J. Disabil. Res., vol. 3, no. 5, p. 20240060, 2024.

[20] W. Wang et al., "A systematic review of machine learning models for predicting outcomes of stroke with structured data," PloS One, vol. 15, no. 6, p. e0234722, 2020.

[21] J. Yu, S. Park, S. H. Kwon, C. M. B. Ho, C. S. Pyo, and H. Lee, "AI-based stroke disease prediction system using real-time electromyography signals," Appl. Sci., vol. 10, no. 19, p. 6791, 2020.

1st

ORIGINALITY REPORT

23%	21%	10%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

16%

★ dspace.daffodilvarsity.edu.bd:8080

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

1st

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21