MULTI-MODAL DATA ANALYSIS FOR STROKE PREDICTION: UNVEILING HIDDEN BIOMARKERS THROUGH MACHINE LEARNING

[1*]Omeye, Emmanuel C., [2]Anyaragbu Hope U.
[1,2]Department of Computer Science, Tansian University, Umunya
Email:[1]emmanuel.omeye@tansianuniversity.edu.ng,[2]anyaragbu.hope@tansianuniversity.edu.ng

Article Info
Received: 4/10/ 2024
Revised: 5/11/2024
Accepted 28/12/2024
Corresponding Author's Email: emmanuel.omeye@tansianuniversity.edu.ng
Corresponding Author's Tel: +234 8064116676

ABSTRACT

Stroke, a leading cause of disability and mortality worldwide, presents a critical challenge for healthcare systems. While demographic factors have traditionally guided stroke risk assessment, recent advancements in machine learning and multimodal data analysis offer promising avenues for uncovering hidden biomarkers. In this project, we propose a novel approach that transcends conventional demographic-based models by integrating diverse data modalities, including genetic, imaging, clinical, and lifestyle factors. The multifaceted nature of stroke demands a comprehensive understanding of its underlying mechanisms, which extend beyond simple demographic variables. By harnessing the power of multimodal analysis, our methodology aims to unveil intricate patterns and interactions among these diverse data sources. Through sophisticated machine learning algorithms, we seek to identify subtle yet significant relationships between genetic predispositions, imaging biomarkers, clinical parameters, and lifestyle habits, collectively contributing to stroke risk. Central to our approach is the recognition that stroke is a complex, multifactorial disease influenced by a myriad of interconnected factors. Conventional models often overlook this complexity, relying solely on demographic characteristics such as age, sex, and ethnicity. In contrast, our methodology embraces the richness of multimodal data, enabling the discovery of novel biomarkers that may have been previously obscured. Furthermore, our research extends beyond mere prediction by aiming to elucidate the underlying biological mechanisms driving stroke susceptibility. By unraveling these hidden biomarkers, we can not only enhance the accuracy of predictive models but also gain insights into the patho-physiology of stroke, thus paving the way for more targeted interventions and personalized treatment strategies.

Keywords: Stroke Prediction, Multimodal Analysis, Biomarkers, Machine Learning, Demographics, Precision Medicine, Risk Assessment, Genetic Factors, Imaging Biomarkers, Clinical Parameters, Lifestyle Factors, Personalized Intervention, Predictive Models, Pathophysiology, Precision Medicine.

1. INTRODUCTION

Stroke remains a formidable global health challenge, exerting a significant burden on healthcare systems and profoundly impacting individuals' lives (Mostarina *et al.*, 2024). As one of the leading causes of disability and mortality worldwide, its prevention and early intervention are paramount. Traditionally, stroke risk assessment has heavily relied on demographic factors such as age, sex, and ethnicity. While undeniably important, these demographic variables only scratch the surface of stroke predisposition, often overlooking the complex interplay of genetic, environmental, and lifestyle factors that contribute to an individual's susceptibility. Recognizing the limitations of demographic-based models, there has been a burgeoning interest in leveraging advanced computational techniques, particularly machine learning, to delve deeper into the underlying mechanisms of stroke. This paradigm shift heralds a new era in stroke prediction and prevention—one that transcends the boundaries of conventional wisdom and embraces the multifaceted nature of the disease (Dongchen *et al.*, 2024). At the forefront of this movement is the emerging field of multimodal analysis, which integrates diverse data modalities, including

genetic information, imaging biomarkers, clinical parameters, and lifestyle habits, to uncover hidden patterns and biomarkers that elude traditional approaches. By harnessing the power of multimodal analysis, researchers aim to unlock the intricate relationships between these disparate data sources, unveiling novel insights into stroke pathophysiology and risk stratification (Natasha, F. and Ramakrishnan, K., 2024).

The project titled "Multi-Modal Data Analysis For Stroke Prediction: Unveiling Hidden Biomarkers Through Machine Learning" represents a pioneering effort in this domain. It seeks to push the boundaries of stroke prediction by integrating cutting-edge machine learning algorithms with a comprehensive array of data modalities. By adopting a holistic approach that goes beyond demographics, the project endeavors to identify previously unrecognized biomarkers that hold promise for revolutionizing stroke risk assessment and personalized intervention strategies. Central to the project's objectives is the recognition that stroke is not a monolithic entity but rather a heterogeneous syndrome with diverse etiologies and risk factors. Therefore, a one-size-fits-all approach to risk assessment is inadequate. Instead, by embracing the complexity of stroke etiology and leveraging multimodal analysis, the project aims to develop more accurate predictive models that account for the unique interplay of genetic predispositions, environmental exposures, and lifestyle habits in each individual.

Through this multidisciplinary endeavor, the project aspires to propel the field of stroke research forward, laying the groundwork for a new era of precision medicine in stroke prevention and management. By unraveling the intricate tapestry of stroke predisposition, the project holds the promise of not only enhancing clinical decision-making but also empowering individuals to take proactive steps towards mitigating their stroke risk.

The paper is organized as follows: Section 2 presents an overview of related works. Subsequently, in section 3, we show the **system design and implementation.** Finally, in section 4, we present the **conclusion and future direction**.

2. RELATED WORKS

Chowdhary Hassan Raza (2024). Stroke Data Analysis and Prediction

This paper explores the impact of strokes on society and emphasizes collaborative efforts to enhance stroke management. By leveraging technology and medical records, caregivers gain insights into relevant risk factors for stroke prediction. This paper systematically examines various components in electronic health records for effective stroke forecasting. Employing diverse statistical methods and principal component analysis, we identify the most significant factors for stroke prediction. Using statistical methods and principal component analysis, we identify age, heart disease, average glucose level, and hypertension as crucial factors. Our findings indicate that age, heart disease, average glucose level, and hypertension emerge as the most critical factors for detecting stroke in patients. Furthermore, a perceptron neural network utilizing these attributes achieves superior accuracy and lower miss rates compared to other methods on a balanced dataset. Keywords: Age, heart disease, glucose level, hypertension.

Hewei, W., Chidozie, S., Nishtha, J., Bharadwaj, V., and Deepu, J. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks.
The negative impact of stroke in society has led to concerted efforts to improve the management and diagnosis of stroke. With an increased synergy between technology and medical diagnosis, caregivers create opportunities for better patient management by systematically mining and archiving the patients' medical records. Therefore, it is vital to study the interdependency of these risk factors in patients' health records and understand their relative contribution to stroke prediction. This paper systematically analyzes the various factors in electronic health records for effective stroke prediction. Using various statistical techniques and principal component analysis, we identify the most important factors for stroke prediction. We conclude that age, heart disease, average glucose level, and hypertension are the most important factors for detecting stroke in patients. Furthermore, a perceptron neural network using these four attributes provides the highest accuracy rate and lowest miss rate compared to using all available input features and other benchmarking algorithms. As the dataset is highly imbalanced concerning the occurrence of stroke, we report our results on a balanced dataset created via sub-sampling techniques.

Elias, D. and Maria, T. (2022). Stroke Risk Prediction with Machine Learning Techniques.
This study investigated the potential of machine learning (ML) to predict stroke risk. Researchers developed and compared several models to create a reliable system for long-term stroke prediction. Their key contribution was a stacking method that achieved impressive results, validated by metrics like AUC (Area Under the Curve), precision, recall, F-measure, and accuracy. This method outperformed others, reaching an AUC of 98.9%, F-measure, precision, and recall of 97.4%, and an accuracy of 98%.

Nojood, A., Rahaf, A., Rehab, A., and Lubna, A. (2023). Using Machine Learning Algorithm as a Method for Improving Stroke Prediction
Sudden strokes have profoundly impacted society, prompting concerted efforts to enhance stroke diagnosis and management. Technological advancements have revolutionized the medical field, providing caregivers with robust tools to mine and archive patients' medical records for efficient retrieval. Understanding the risk factors predisposing individuals to strokes is paramount, facilitating more accurate prediction and preventive measures.
This research delves into the factors influencing stroke prediction processes, leveraging electronic health records for analysis. Through statistical methods and Principal Component Analysis (PCA), the study identifies key variables crucial for accurate stroke prediction. Age, average glucose level, heart disease, and hypertension emerge as pivotal factors affecting stroke prognosis. To address the challenge of imbalanced datasets, a balanced dataset is created through sub-sampling, ensuring model efficacy during evaluation. Seven machine learning algorithms—Naïve Bayes, SVM, Random Forest, KNN, Decision Tree, Stacking, and majority voting—are implemented on the Kaggle dataset to predict stroke occurrences.
Following dataset preprocessing and division into training and testing subsets, the proposed algorithms undergo evaluation based on metrics including accuracy, f1 score, recall, and

precision. The Naïve Bayes classifier exhibits the lowest accuracy (86%), whereas other algorithms achieve comparable accuracies of 96%, f1 scores of 0.98, precision of 0.97, and recall of 1.

This study underscores the importance of leveraging electronic health records and advanced machine learning techniques for stroke prediction. By identifying critical risk factors and evaluating predictive models, it contributes to enhancing stroke prognosis accuracy, ultimately facilitating timely interventions and improved patient outcomes.

Krishna, M., Sandesh, G., Jungpil, S., Anmol, A., Md. Mezbah, U., M. Firoz M. (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention.

Stroke, a perilous medical condition, arises from the disruption of blood flow to the brain, leading to neurological impairment. Its global prevalence poses significant health and economic challenges, prompting researchers to develop automated stroke prediction algorithms to enable timely interventions and potentially save lives. With the aging population increasing the number of individuals at risk, the need for precise and effective prediction systems has never been more pressing. In this study, we conducted a comparative analysis of six well-known classifiers to evaluate the efficacy of the proposed machine learning (ML) technique in terms of generalization capability and prediction accuracy. Additionally, we explored two explainable techniques—SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations)—to shed light on the decision-making process of black-box ML models, especially pertinent in the medical domain. SHAP and LIME have emerged as reliable methodologies for elucidating model decisions, offering insights into complex model behaviors. By integrating these explainable techniques into our analysis, we aimed to enhance the interpretability of the ML models employed for stroke prediction.

The experimental findings unveiled the superiority of more intricate models over simpler ones, with the top-performing model achieving an impressive accuracy rate of nearly 91%, while other models achieved accuracies ranging from 83% to 91%. This underscores the potential of advanced ML techniques in bolstering stroke prediction accuracy and guiding clinical decision-making. The proposed framework, encompassing both global and local explainable methodologies, holds promise in standardizing complex ML models and unraveling their decision-making rationale. This deeper understanding can inform healthcare professionals in devising more effective stroke care and treatment strategies, ultimately improving patient outcomes and mitigating the burden of stroke on society.

Dongchen, W., Xinfang, Z. and Xiaochen, Z. (2024). A machine learning-based model for stroke prediction.

“A machine learning-based model for stroke prediction" explores the effectiveness of various machine learning techniques in predicting strokes. The study highlights the importance of early stroke prediction for improving treatment outcomes and intervention strategies. By analyzing a dataset comprising 5110 records with 12 attributes, the authors employed several machine learning models to assess their predictive performance.

The models evaluated include Logistic Regression, Support Vector Classifier (SVC), K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier (XGBClassifier), and Deep Neural Networks (DNN). After preprocessing the data to remove redundancies and address dataset imbalance, the researchers compared these models. The findings revealed that the Deep Neural Networks (DNN) model performed the best, achieving an AUC of 82% in scenarios with extreme data imbalance, surpassing the performance of the other models. The paper underscores the significance of model choice in stroke prediction and provides insights into how machine learning can enhance early diagnosis and intervention strategies.

Chunhua, G. and Wang, H. (2024). Intelligent Stroke Disease Prediction Model Using Deep Learning Approaches.

"Intelligent Stroke Disease Prediction Model Using Deep Learning Approaches," address the critical need for early stroke detection to improve intervention outcomes. They highlight the severe health risks associated with stroke, emphasizing the importance of recognizing warning signs early. Their study leverages deep learning techniques, specifically deep neural networks, to enhance stroke prediction.

The researchers propose a novel model that integrates various physiological parameters with advanced deep learning frameworks, including Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP) and regression networks. To tackle the issue of imbalance in stroke datasets, they employ data augmentation techniques and WGAN-GP to generate high-fidelity stroke data for training their prediction model. This approach allows the model to better handle the disparity between positive and negative samples.

Their model represents the relationship between physiological characteristics and stroke risk through a nonlinear mapping transformation, resulting in a deep regression network specifically designed for stroke prediction. When compared to traditional machine learning algorithms such as decision trees, random forests, support vector machines, and artificial neural networks, the proposed deep learning method demonstrates superior performance, particularly in comprehensive measurement indices. The study's results suggest that their model not only performs optimally in prediction but also exhibits robustness, making it a valuable tool for stroke disease prediction.

3. SYSTEM DESIGN AND IMPLEMENTATION

3.1 Objectives of the Design

Developing and establishing a Diagnostic and Treatment Recommendation System for stroke are guided by a comprehensive set of objectives. These objectives are strategically aligned to craft a powerful and efficient solution tailored to meet the unique requirements of healthcare professionals and individuals in managing stroke effectively. The overarching objectives encompass both the technical dimensions of system design and the strategic goals associated with optimizing healthcare analytics. Here is a detailed overview:

i. **Accurate Diagnosis:** The primary objective of the design is to develop a diagnostic system that can accurately identify and diagnose stroke in individuals. The system should utilize advanced algorithms and machine learning techniques to analyze relevant medical data, such as blood pressure readings, medical history, and lifestyle factors, to provide a precise and reliable diagnosis.

ii. **Risk Stratification:** Implement a risk stratification component within the system to categorize patients based on the severity of their stroke and associated risk factors. This will aid healthcare professionals in prioritizing and customizing treatment plans according to the individual patient's needs, ensuring timely intervention for those at higher risk.

iii. **Personalized Treatment Recommendations**: Design the system to generate personalized treatment recommendations for individuals diagnosed with stroke. The recommendations should take into account factors such as age, gender, comorbidities, lifestyle choices, and medication adherence. This personalized approach aims to enhance treatment efficacy and patient compliance.

iv. **Integration of Patient Data**: Develop a user-friendly interface that allows seamless integration of patient data from various sources, including electronic health records (EHRs), wearable devices, and self-reported information. This integration ensures a comprehensive view of the patient's health, facilitating more informed decision-making by healthcare professionals.

v. **Continuous Monitoring and Feedback**: Implement a monitoring system that provides continuous feedback on the patient's health status. The design should enable real-time tracking of blood pressure trends, medication adherence, and lifestyle changes. This feature empowers both patients and healthcare providers to make timely adjustments to the treatment plan, fostering better long-term management of stroke.

vi. **Decision Support for Healthcare Professionals**: Create a decision support system to assist healthcare professionals in making informed decisions about treatment strategies. The system should present relevant clinical guidelines, research findings, and treatment options, aiding physicians in tailoring interventions based on the latest evidence and individual patient characteristics.

3.2 Submenus / Subsystems

These are menus or headings that provide a structured interface for navigating through key aspects of the project.

3.2.1 Stroke Diagnostic System

Within the main framework of the Stroke Prediction System, the Diabetes Diagnostic Submodule is tailored to tackle the intersection of diabetes within the stroke population. This segment seamlessly incorporates blood glucose monitoring data into the diagnostic workflow, enhancing the assessment of diabetes prevalence among stroke patients. Leveraging specialized risk assessment tools, it evaluates the likelihood of diabetes co-occurrence by analyzing patient

profiles, medical backgrounds, and pertinent clinical markers. Distinguished algorithms dedicated to diabetes diagnosis ensure precise detection and categorization of diabetes cases within the stroke cohort. Moreover, the submodule emphasizes the delivery of personalized treatment recommendations, taking into account both hypertension and diabetes. By offering a holistic approach to managing these overlapping conditions, it strives to optimize patient care and outcomes in the context of stroke prevention and treatment.

3.2.2 Performance and Evaluation Analysis Module

The System Performance Evaluation submodule is dedicated to scrutinizing and refining the overall efficacy of the system. Central to its mission is the examination of healthcare professionals' usage patterns within the system, with a focus on identifying frequently accessed features and potential areas for enhancement. Through meticulous tracking and analysis, it aims to optimize system usability and efficiency. This submodule delves into the speed and effectiveness of data processing algorithms, prioritizing the timely delivery of diagnostic outcomes and treatment suggestions. By scrutinizing algorithmic performance, it ensures swift and accurate responses to user queries, fostering streamlined patient care. Furthermore, the submodule integrates mechanisms for gathering user feedback, providing valuable insights into system responsiveness and user satisfaction. This feedback loop facilitates ongoing refinement, driving continuous improvement efforts.

Understanding and analyzing these variables collectively contribute to the creation of a robust diagnostic and recommendation system, enabling a comprehensive approach to hypertension management and patient care.

The dataset is formatted as a CSV file and can be downloaded from Kaggle. It includes a readme file that provides more information about the data collection process and the format of the data. To analyze the data, I first utilized Pandas' read_csv() function to import the relevant information from the provided CSV file. Leveraging the power of Pandas, I was able to efficiently import the dataset stored in a CSV file using the read_csv() method.

Index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
9374	41	1	2	115	564	0	0	160	0	1.6	1	0	3	1
140	49	0	2	120	295	0	0	157	0	0.6	2	0	2	1
19176	72	1	2	140	197	0	2	116	0	1.1	1	0	2	1
1573	61	1	2	94	199	0	1	179	0	0	2	0	2	1
23819	72	1	2	112	230	0	0	165	0	2.5	1	1	3	0
20621	79	0	0	146	218	0	1	105	0	2	1	1	3	0
16847	42	1	0	138	294	1	1	106	0	1.9	1	3	2	0
22905	59	0	3	170	288	0	0	159	0	0.2	1	0	3	0
8200	67	0	0	144	200	0	0	126	1	0.9	1	0	3	0
23359	83	0	0	145	174	0	1	125	1	2.6	0	0	3	0
10118	50	1	1	134	201	0	1	158	0	0.8	2	1	2	1
945	61	0	0	122	222	0	0	186	0	0	2	0	2	1

Figure 1: A section of the dataset

The dataset underwent feature scaling using the **StandardScaler** module from the **sklearn.preprocessing** library. This preprocessing step standardized the numerical features, aligning them with a mean of 0 and a standard deviation of 1. The utilization of **StandardScaler** aimed to normalize feature values, addressing varying scales and optimizing the dataset for subsequent stages of the project. The figure below illustrates this.

	0	1	2	3	4
0	-0.172579	1.00029	-0.934922	-0.0925932	-2.22607
1	0.0250976	1.00029	-0.934922	0.0206979	0.00751971
2	0.552235	-0.999712	-0.934922	-0.545758	0.758469
3	-0.699716	-0.999712	-0.934922	-0.659049	-1.12853
4	-1.02918	-0.999712	-0.934922	-1.56538	-0.743429
5	1.93597	-0.999712	0.0432152	-0.432467	0.277091
6	0.815804	1.00029	-0.934922	0.0206979	0.00751971
7	-0.238471	-0.999712	-0.934922	-0.659049	-0.935981
8	0.684019	-0.999712	1.02135	-1.2255	-1.37885
9	-0.0407946	1.00029	-0.934922	-0.659049	-1.12853

Figure 2: A result of the StandardScaler on the dataset

3.3 Algorithm

In the study on stroke prediction, ensemble methods were a key strategy to enhance model performance. Specifically, the stacking ensemble technique was employed. Stacking, or stacked generalization, involves combining multiple base models to create a more robust and accurate predictive system. In this approach, various machine learning algorithms are trained on the same dataset, and their predictions are then used as inputs for a meta-model. This meta-model learns to make the final prediction by integrating the outputs from the base models.

By incorporating stacking, the study was able to leverage the unique strengths of each base model while minimizing their individual weaknesses. This approach likely improved the handling of complex data patterns and led to more reliable predictions. The use of stacking contributed to a more effective early prediction tool for stroke, showcasing the power of ensemble methods in achieving better performance and accuracy. Let's explore some other the key models integral to the success of this innovative approach.

3.4 The Correlation Matrix

In data analytics, the corr() method is a fundamental tool employed to calculate the correlation matrix, revealing the relationships between variables in a dataset. Correlation measures the statistical association between two or more variables, indicating the degree to which changes in one variable correspond to changes in another. The corr() method is typically applied to a

DataFrame in Python, often using libraries such as Pandas. The corr() method operates on a Pandas DataFrame, with each column representing a different variable. It computes pairwise correlation coefficients for all variable combinations in the DataFrame.

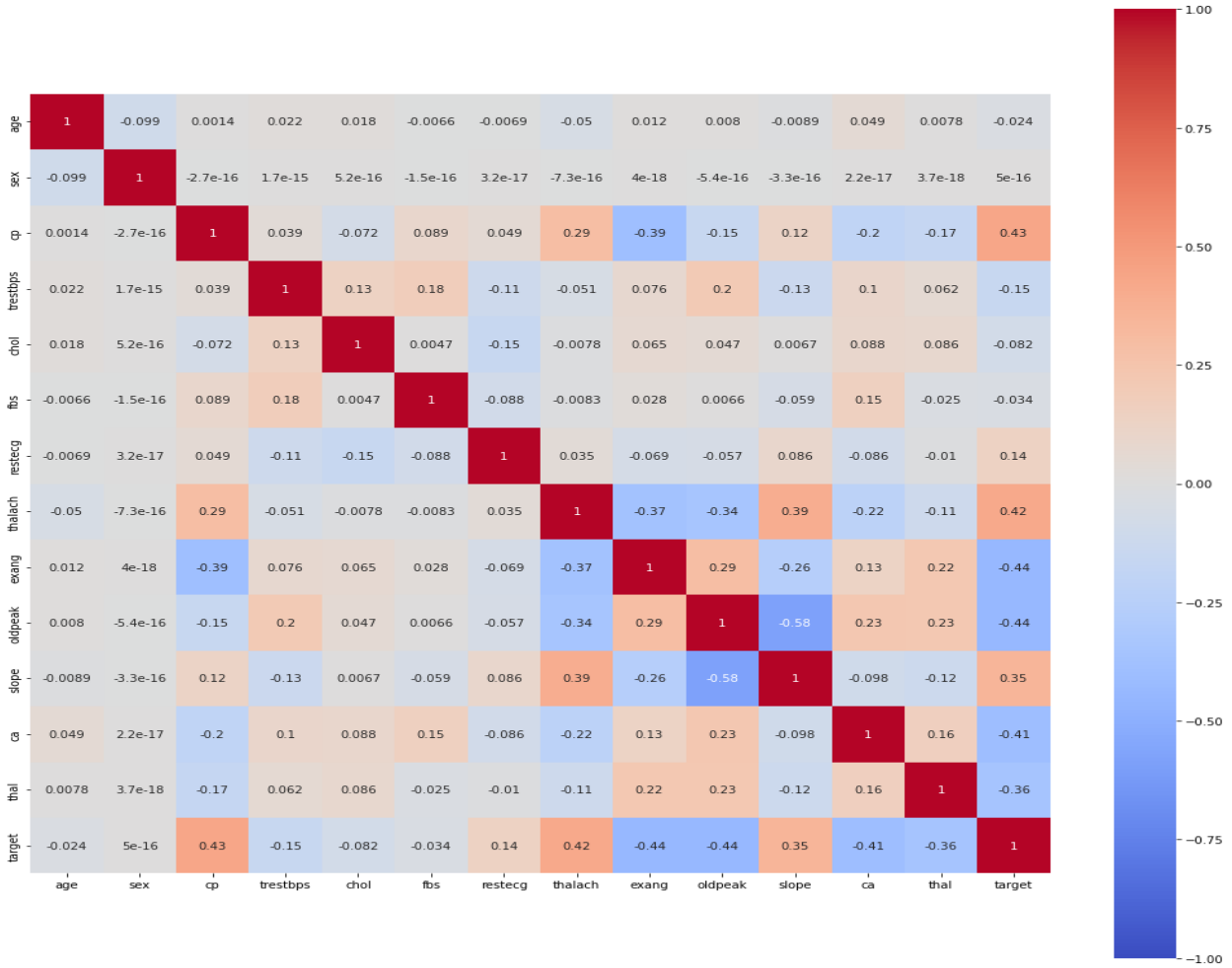

Figure 3: Figure showing the correlation matrix gotten during analysis.

By applying the corr() method, data analysts can gain valuable insights into the interdependencies among variables, aiding in feature selection, identifying multicollinearity, and informing subsequent analyses in various data science and machine learning tasks.

3.5 Confusion Matrix

A confusion matrix is a fundamental tool in evaluating the performance of a classification model. It provides a comprehensive and detailed summary of the model's predictions, breaking down the outcomes into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These components are crucial for assessing the effectiveness of a model across various metrics. The following are the components of the Confusion Matrix:

i. **True Positive (TP):** Instances where the model correctly predicts the positive class. For example, correctly identifying patients with hypertension.

ii. **True Negative (TN):** Instances where the model correctly predicts the negative class. For example, correctly identifying patients without hypertension.

iii. **False Positive (FP)**: Instances where the model predicts the positive class incorrectly. Also known as a Type I error. For example, predicting hypertension in a patient who does not have it.

iv. **False Negative (FN):** Instances where the model predicts the negative class incorrectly. Also known as a Type II error. For example, failing to predict hypertension in a patient who actually has it.

3.6 The Logistic Regression Model

Logistic Regression is a statistical method commonly used in data analytics and machine learning for binary classification problems. Despite its name, it is employed for predicting the probability of an instance belonging to a particular category rather than predicting a continuous outcome. Logistic Regression is well-suited for problems where the dependent variable is binary, meaning it has only two possible outcomes (e.g., 0 or 1, True or False). It applies the sigmoid (logistic) function to transform a linear combination of input features into a value between 0 and 1. This transformed value represents the probability of belonging to the positive class.

Logistic Regression establishes a decision boundary based on the calculated probabilities. For binary classification, a common threshold is set at 0.5. If the predicted probability is above 0.5, the instance is classified as the positive class; otherwise, it belongs to the negative class. From the above Confusion matrix evaluation of the Logistic Regression model, the accuracy of 0.8619, Precision score of 0.8652, recall score of 0.8619, and F1 score value of 0. 0.8606 was achieved.

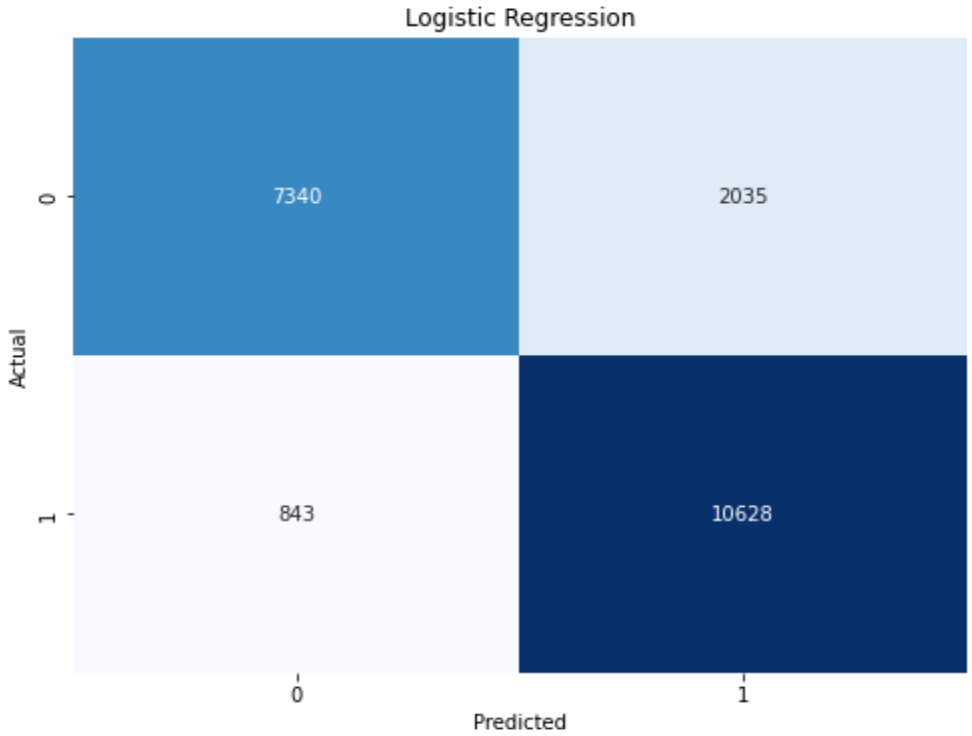

Figure 4: Figure showing the confusion matrix of the Logistic Regression model

3.7 K Nearest Neighbor

K-Nearest Neighbors (KNN) is a versatile and straightforward algorithm used for both classification and regression tasks in data analytics and machine learning. It is a non-parametric, instance-based learning algorithm, meaning it doesn't make assumptions about the underlying data distribution and makes predictions based on the similarity of new instances to existing data points. It makes predictions based on the majority class or average value of the k-nearest data points to a given instance. The value of k is a user-defined parameter. KNN relies on a distance metric (commonly Euclidean distance) to measure the similarity between instances. The algorithm identifies the k-nearest neighbors by finding the data points with the smallest distances to the target instance. KNN can be computationally expensive, especially with large datasets, as it requires calculating distances for each prediction. From the Confusion matrix evaluation of the KNN model, the accuracy of 0.1, Precision score of 0.1, recall score of 0.1and F1 score value of 0.1was achieved.

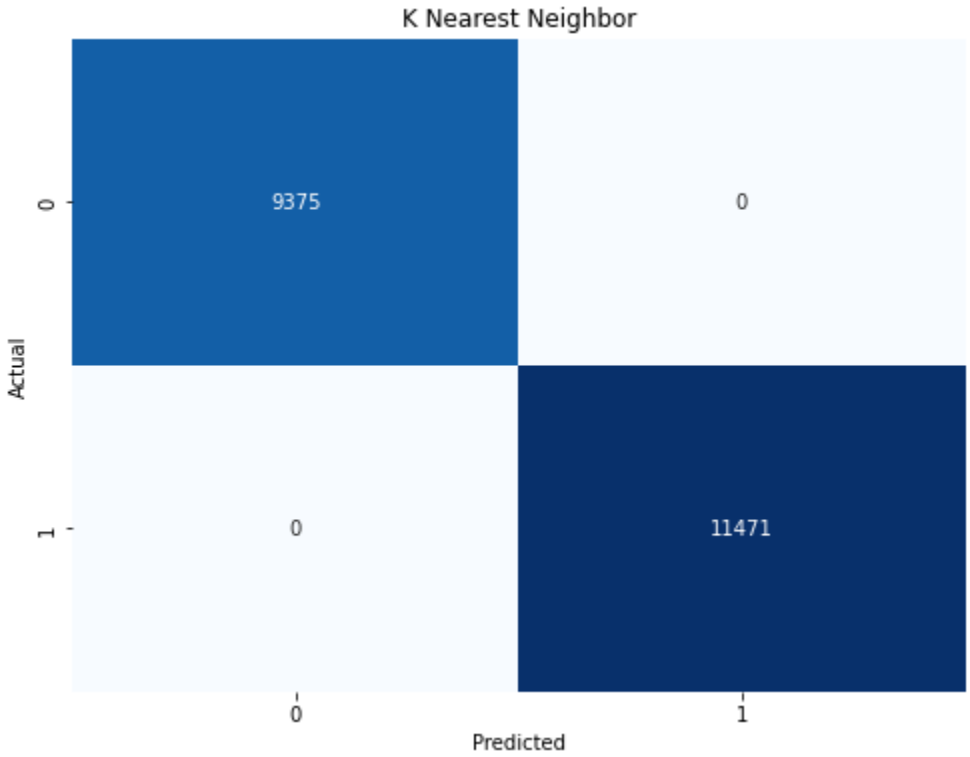

Figure 5: Confusion matrix showing the perfect performance of the KNN model

3.8 Decision Tree

A Decision Tree is a powerful and interpretable machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a decision based on a particular feature, each branch represents the outcome of the decision, and each leaf node represents the final decision or the predicted output. Decision Trees are widely used due to their simplicity, ability to handle both numerical and categorical data, and their capability to capture complex relationships in the data. Decision Trees have a hierarchical

structure with nodes representing decisions or tests based on specific features. The tree structure flows from the root node to the leaf nodes, where the final decisions or predictions are made. Decision Trees employ various splitting criteria, such as Gini impurity or mean squared error, to determine how well a particular feature separates the data into distinct classes or groups. From the Confusion matrix evaluation of the Decision Tree model given below, the accuracy of 0.1, Precision score of 0.1, recall score of 0.1and F1 score value of 0.1was achieved.

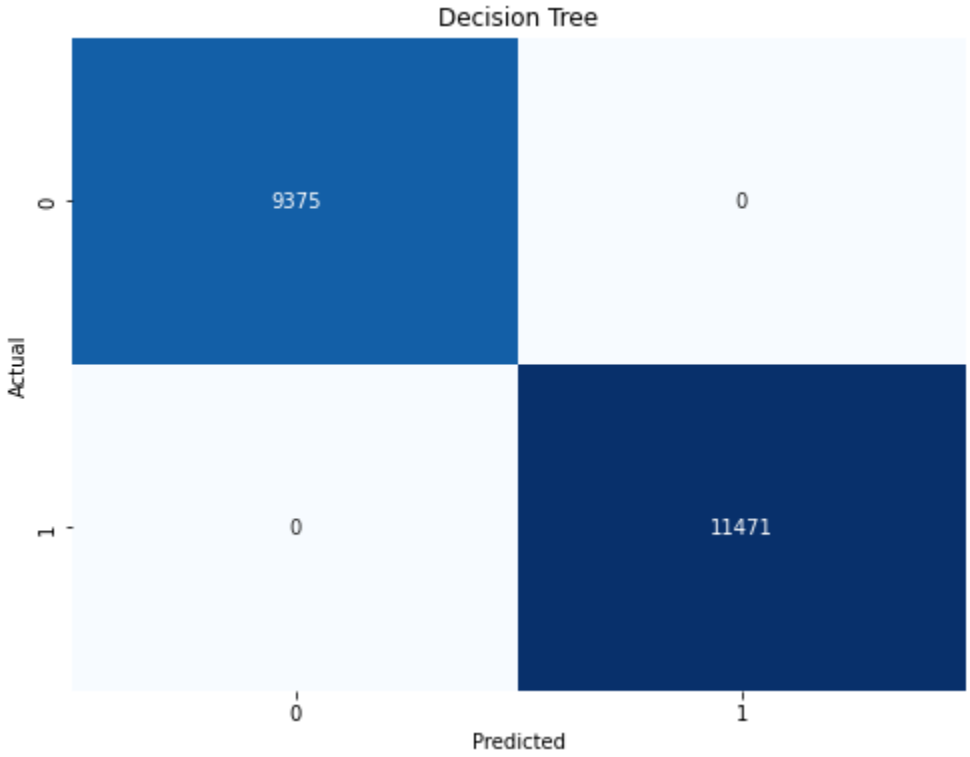

Figure 6: Confusion Matrix showing the perfect performance of the Decision Tree model.

3.9 Random Forest

For this study the Random Forest model was used as the META model of the stack ensembled system, which comprises of the three aforementioned algorithms: Logistic Regression, K Nearest Neighbor and Decision Tree. Random Forest is a potent ensemble learning algorithm applicable to both classification and regression tasks. It constructs multiple decision trees during training and combines their predictions, providing a robust and versatile solution in data analytics and machine learning. Random Forest builds multiple decision trees and aggregates their predictions to enhance accuracy and stability. From the Confusion matrix evaluation of the Random Forest model given below, the accuracy of 0.1, Precision score of 0.1, recall score of 0.1and F1 score value of 0.1 was achieved.

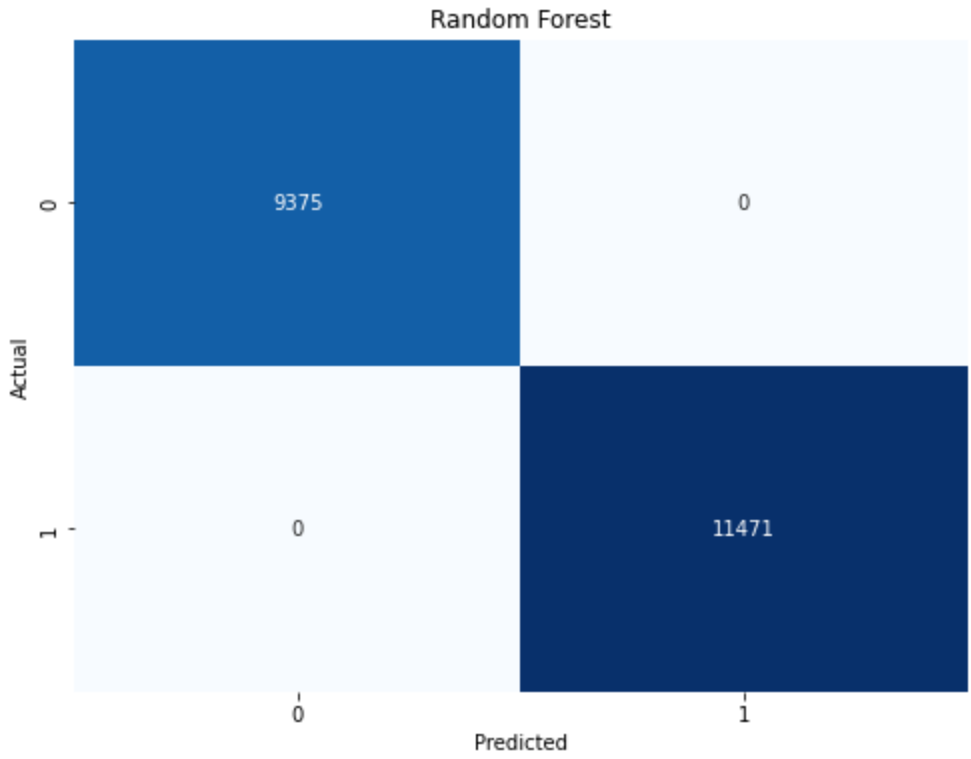

Figure 7: Confusion matrix showing the perfect performance of the Random Forest Model.

The algorithm employs bagging (Bootstrap Aggregating), training each tree on a random subset of the dataset sampled with replacement. Random Forest introduces randomness by considering a random subset of features at each decision tree split, reducing correlation between trees. For classification tasks, the final prediction is determined by majority vote, and for regression, it is based on the average prediction. The table 4.1 below shows the summary of the evaluation of the models in the stack system.

Table 1: Summary of the evaluation of the models in the Stack Ensembled system.

Models	Accuracy	Precision	Recall	F1 score
K-NN	1.0000	1.0000	1.0000	1.0000
Logistic Regression	0.8619	0.8652	0.8619	0.8606
Decision Tree	1.0000	1.0000	1.0000	1.0000
Random Forest (Meta)	1.0000	1.0000	1.0000	1.0000

3.10 Evaluation Metrics

As we navigate the landscape of model evaluation in data analytics and machine learning, it becomes crucial to employ metrics that provide nuanced insights into the performance of our models. Among these, the Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve stand as invaluable tools. These metrics go beyond simple accuracy, offering a deeper understanding of a model's ability to discriminate between classes and balance precision and

recall. In the upcoming discussion, we delve into the significance and interpretation of the ROC curve and Precision-Recall curve, unraveling their role in assessing the efficacy and reliability of our predictive models.

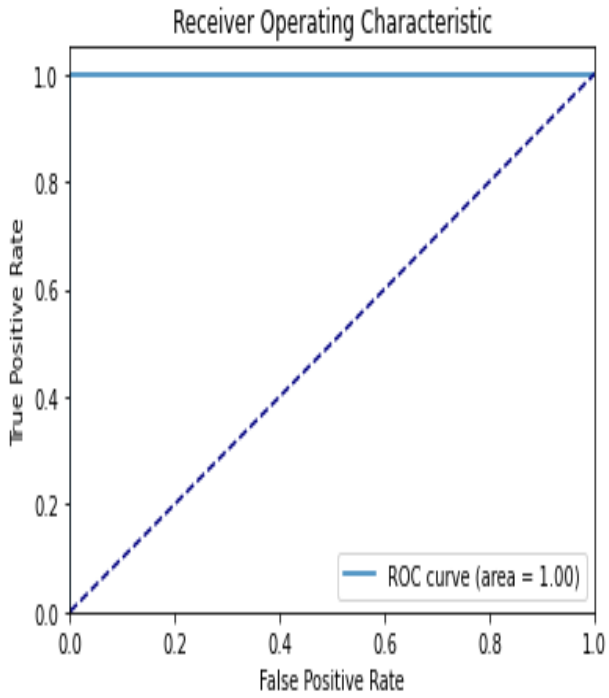

Figure 8: Random Forest ROC model

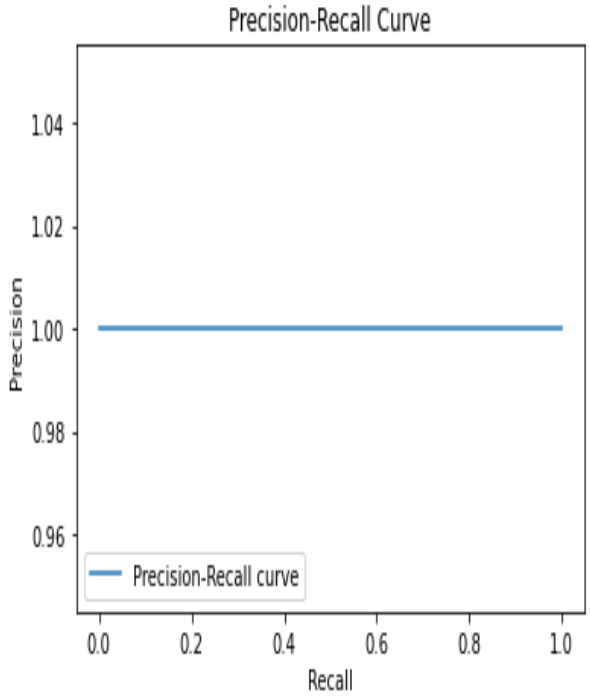

Figure 9: Precision-Recall of the Random Forest

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between true positive rate (sensitivity) and false positive rate across various classification thresholds. A perfect ROC curve would be one where the true positive rate is always 1 (100%), and the false positive rate is always 0 (0%). A perfect ROC curve starts from the bottom-left corner and ascends vertically to the top-left corner. This indicates that the model is achieving a true positive rate of 1 without incurring any false positives just like the one we have in Figure 4.8.

The Precision-Recall (PR) curve is a valuable metric for evaluating classification models, particularly in scenarios where class imbalances exist. It showcases the trade-off between precision and recall at various decision thresholds. A perfect PR curve indicates a model with flawless precision and recall across all thresholds. A perfect PR curve starts from the bottom-right corner and ascends vertically to the top-right corner. This signifies that the model achieves both perfect precision and recall. At the top-right corner, precision is maximized, indicating that every positive prediction made by the model is indeed correct. Simultaneously, recall is maximized, implying that the model captures every positive instance in the dataset.

The area under the Precision-Recall curve (AUC-PR) quantifies the overall performance of the model. In the case of a perfect curve, the AUC-PR value would be 1, reflecting flawless precision and recall trade-offs. Similar to the ROC curve interpretation, a random classifier would follow the 45-degree diagonal line (the line of no-discrimination) from the bottom-right to

the top-left. A perfect model deviates significantly from this diagonal, reaching the top-right corner.

4. CONCLUSION AND FUTURE DIRECTION

In conclusion, the Stroke Prediction System represents a significant advancement in the field of stroke prevention and management. By integrating advanced machine learning techniques with comprehensive data analysis, this project has laid the groundwork for more accurate and personalized stroke risk assessment. Through the identification of key biomarkers and the development of predictive models, the system has the potential to revolutionize clinical decision-making and improve patient outcomes. The findings from this project underscore the importance of moving beyond traditional demographic-based models and embracing a multimodal approach to stroke prediction. By considering a wide range of factors, including genetic, imaging, clinical, and lifestyle data, the system has provided valuable insights into the complex interplay of variables that influence stroke susceptibility.

Moreover, the evaluation of various machine learning algorithms has highlighted the importance of selecting appropriate models for stroke prediction. While some models may exhibit higher accuracy than others, the choice of algorithm should be guided by considerations of interpretability, generalization capability, and clinical utility.

4.2 Future Directions

Looking ahead, there are several avenues for further research and development in the field of stroke prediction:

i. **Refinement of Predictive Models**: Continued refinement and optimization of predictive models will be essential to enhance their accuracy and reliability. This may involve incorporating additional data sources, refining feature selection techniques, and exploring novel algorithmic approaches.

ii. **Integration of Real-Time Data**: The integration of real-time data streams, such as wearable device data and remote monitoring technologies, holds promise for improving the timeliness and accuracy of stroke prediction. By continuously monitoring patients' health status, clinicians can intervene proactively to prevent stroke events.

iii. **Validation and External Validation**: It is crucial to validate the performance of the Stroke Prediction System on diverse and independent datasets to assess its generalizability and robustness across different patient populations and healthcare settings. External validation studies will provide valuable insights into the system's real-world applicability and performance.

iv. **Clinical Implementation and Decision Support**: The ultimate goal of the Stroke Prediction System is to translate research findings into clinical practice. Future efforts should focus on integrating the system into existing healthcare workflows and providing decision support tools to assist clinicians in interpreting predictive results and making informed treatment decisions.

v. **Longitudinal Studies and Outcome Prediction**: Longitudinal studies are needed to investigate the long-term prognostic value of the Stroke Prediction System and its ability

to predict stroke recurrence, functional outcomes, and mortality. By tracking patients over time, researchers can assess the system's effectiveness in guiding personalized treatment and monitoring disease progression.

In summary, the Stroke Prediction System holds immense potential to transform stroke care by enabling early detection, risk stratification, and personalized intervention. Continued research and innovation will be essential to realize the full promise of this technology and improve outcomes for stroke patients worldwide.

5. REFERENCES

Chowdhary Hassan Raza (2024). Stroke Data Analysis and Prediction. *International Journal of Scientific Research in Engineering and Management 08(03):1-5.*http://dx.doi.org/10.55041/IJSREM29536.

Hewei, W., Chidozie, S., Nishtha, J., Bharadwaj, V., and Deepu, J. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks.

Elias, D. and Maria, T. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors 22(13):4670*. http://dx.doi.org/10.3390/s22134670.

Nojood, A., Rahaf, A., Rehab, A., and Lubna, A. (2023). Using Machine Learning Algorithm as a Method for Improving Stroke Prediction. *International Journal of Advanced Computer Science and Applications 14(4).*http://dx.doi.org/10.14569/IJACSA.2023.0140481.

Krishna, M., Sandesh, G., Jungpil, S., Anmol, A., Md. Mezbah, U., M. Firoz M. (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention. *IEEE Access PP(99):1-1*. *http://dx.doi.org/10.1109/ACCESS.2023.3278273*.

Dongchen, W., Xinfang, Z. and Xiaochen, Z. (2024). A machine learning-based model for stroke prediction. *Applied and Computational Engineering 78(1):122-130*. http://dx.doi.org/10.54254/2755-2721/78/20240645

Chunhua, G. and Wang, H. (2024). Intelligent Stroke Disease Prediction Model Using Deep Learning Approaches. *Stroke Research and Treatment 2024:1-10.*http://dx.doi.org/10.1155/2024/4523388

Natasha, F. and Ramakrishnan, K. (2024). Machine Learning-Driven Stroke Prediction Using Independent Dataset. *JOIV International Journal on Informatics Visualization. 8(2):1030.*http://dx.doi.org/10.62527/joiv.8.2.2689

Mostarina, M., Mahedy, H., Palash, U., and Seifedine, K. (2024). A stroke prediction framework using explainable ensemble learning. Computer Methods in Biomechanics & Bio Engineering. http://dx.doi.org/10.1080/10255842.2024.2316877