Predictive Modeling of Stroke Risk in High-Risk Populations Using Machine Learning Techniques

Thesis Report

CSE 4100

Submitted by

Antika Ghosh	**190104005**
Purna Chandra Saha	**20200104141**
Apu Das	**20200204108**

Supervised by

Dr. S. M. A. AL-Mamun

Professor

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

December 14, 2024

ABSTRACT

Stroke remains one of the leading causes of mortality and long-term disability worldwide. Early prediction of stroke risk can significantly reduce this burden by facilitating timely preventive interventions. This study explores the use of machine learning algorithms to predict stroke risk based on demographic, clinical, and lifestyle factors. A range of algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, AdaBoost, XGBoost, and CatBoost, are evaluated for their effectiveness in stroke risk prediction. To address challenges such as imbalanced datasets and feature redundancy, advanced data preprocessing, feature selection, and synthetic data augmentation techniques are employed. Methods such as SMOTE and Tomek Links are used to handle class imbalance, while Recursive Feature Elimination (RFE) and correlation-based filters are applied for feature selection. The models are assessed using performance metrics including accuracy, precision, recall, and F1-score. Random Forest, XGBoost, and Decision Tree exhibit superior performance, with Random Forest achieving an accuracy of 96.49% and F1-score of 0.9660. These findings highlight the potential of machine learning models in advancing personalized healthcare and improving stroke risk stratification, thereby supporting clinical decision-making for early intervention.

Contents

List of Figures

List of Tables

[hyphens]url hyperref article [utf8]inputenc [hyphens]url hyperref

Chapter 1

Introduction

1.1 Introduction

Stroke, a leading cause of death and disability worldwide, continues to pose a significant public health challenge. Despite advancements in medical science, the incidence and mortality rates associated with stroke remain substantial. Timely identification of individuals at high risk of stroke is crucial for implementing preventive measures and initiating prompt treatment. Traditional risk assessment methods, often relying on a limited number of clinical risk factors, have limitations in accurately predicting stroke occurrence. These methods may overlook subtle patterns and complex interactions between various factors that contribute to stroke risk. To address this challenge, machine learning, a powerful tool for analyzing complex data, offers a promising approach. By leveraging advanced algorithms and large datasets, machine learning models can identify intricate patterns and relationships between numerous risk factors, including demographic, clinical, and lifestyle factors. This enables more precise risk stratification and tailored preventive strategies. As of 2024, machine learning has emerged as a promising tool for improving healthcare outcomes. This paper aims to investigate the potential of machine learning techniques in developing a robust predictive model for stroke risk in high-risk populations. By accurately identifying individuals at elevated risk, we can implement targeted interventions to reduce the burden of stroke and improve public health outcomes.

1.2 Problem Statement

Stroke remains a significant global health burden, leading to substantial morbidity and mortality. Traditional risk assessment methods, often based on a limited number of clinical risk factors, have limitations in accurately predicting stroke occurrence. These methods may overlook complex interactions between various factors that contribute to stroke risk. To address this challenge, machine learning offers a promising approach. By leveraging advanced algorithms and large datasets, machine learning models can identify intricate patterns and relationships between numerous risk factors, including demographic, clinical, and lifestyle factors. However, the accuracy and generalizability of machine learning models for stroke risk prediction are influenced by various factors, such as data quality, and model selection. This paper aims to address the following research questions: 1. Data Quality Challenges: How can we effectively preprocess and handle data quality challenges in diverse datasets to improve the accuracy and robustness of stroke risk prediction models? 2. Model Selection and Hyperparameter Tuning: Which machine learning algorithms and hyperparameter configurations are most suitable for accurate stroke risk prediction in high-risk populations? 3. Model Interpretability: How can we interpret the predictions of machine learning models to gain insights into the underlying mechanisms of stroke risk and identify actionable insights for clinical decision-making? By addressing these research questions, this study seeks to advance the state-of-the-art in stroke risk prediction and contribute to the development of more effective preventive and therapeutic strategies.

1.3 Motivation

Recent advancements in machine learning (ML) have demonstrated remarkable potential in transforming stroke prediction and care. Studies such as "Stroke Risk Prediction Using Machine Learning Algorithms" [1] and "A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks" [2] highlight the significant role of ensemble methods and neural networks in achieving high predictive accuracy. These approaches, leveraging demographic and lifestyle data, have set a foundation for stroke risk modeling. Similarly, works like "Predictive Analysis for Risk of Stroke Using Machine Learning Techniques" [3] and "Stroke Risk Prediction with Machine Learning Techniques" [4] emphasize the robust performance of ensemble methods and neural networks in processing high-dimensional medical datasets.

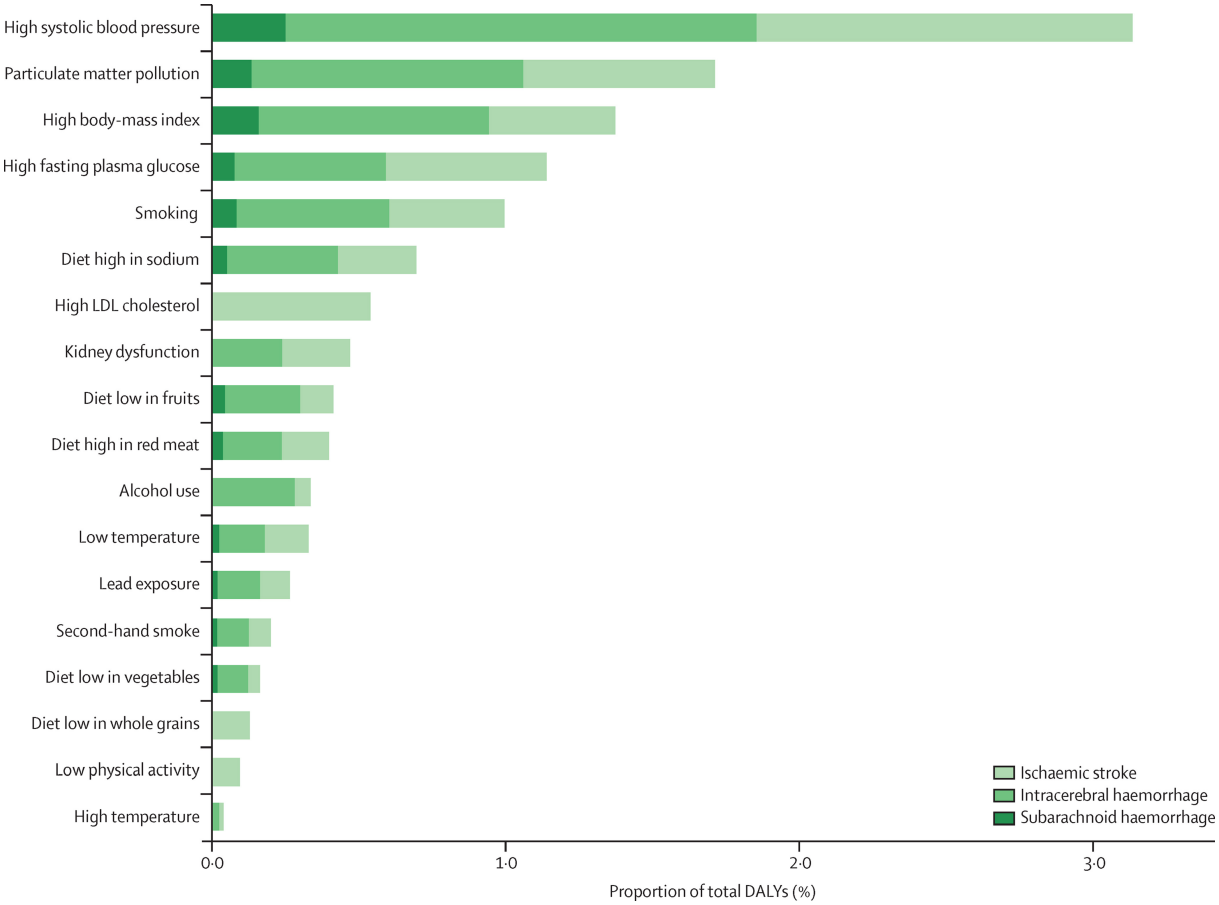

Figure 1.1: Various risk factors for different types of stroke.[18]

Moreover, studies like "Scaling behaviors of deep learning and linear algorithms for the prediction of stroke severity" [8] and "Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review" [7] underscore the transformative potential of deep learning in analyzing MRI-derived lesion data and multimodal data integration, respectively. These approaches have revealed nonlinear relationships and contributed to enhanced prediction accuracy. Despite these advancements, critical limitations remain.

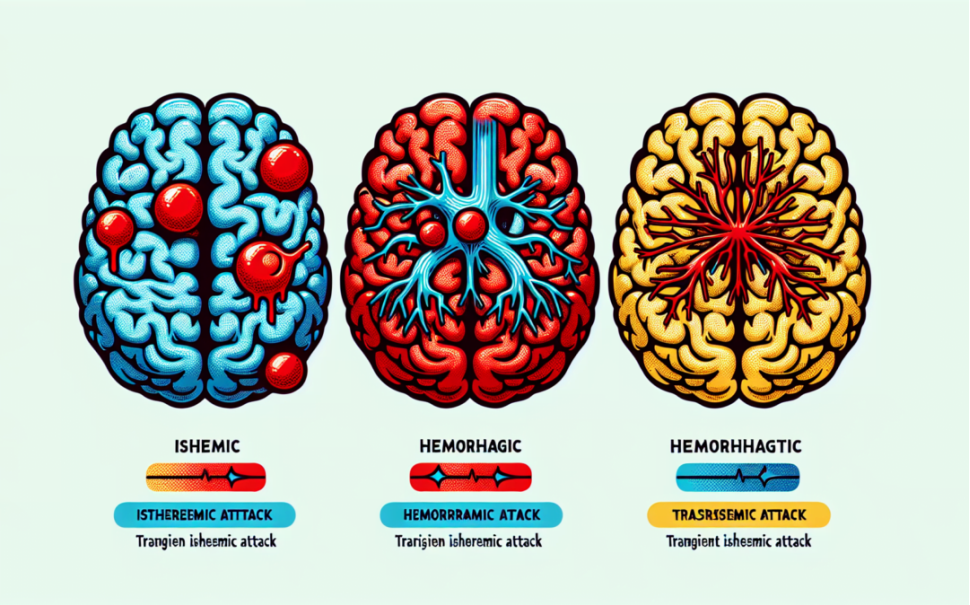

Figure 1.2: Types of stroke: Ischemic, Hemorrhagic, and Transient Ischemic Attack (TIA).

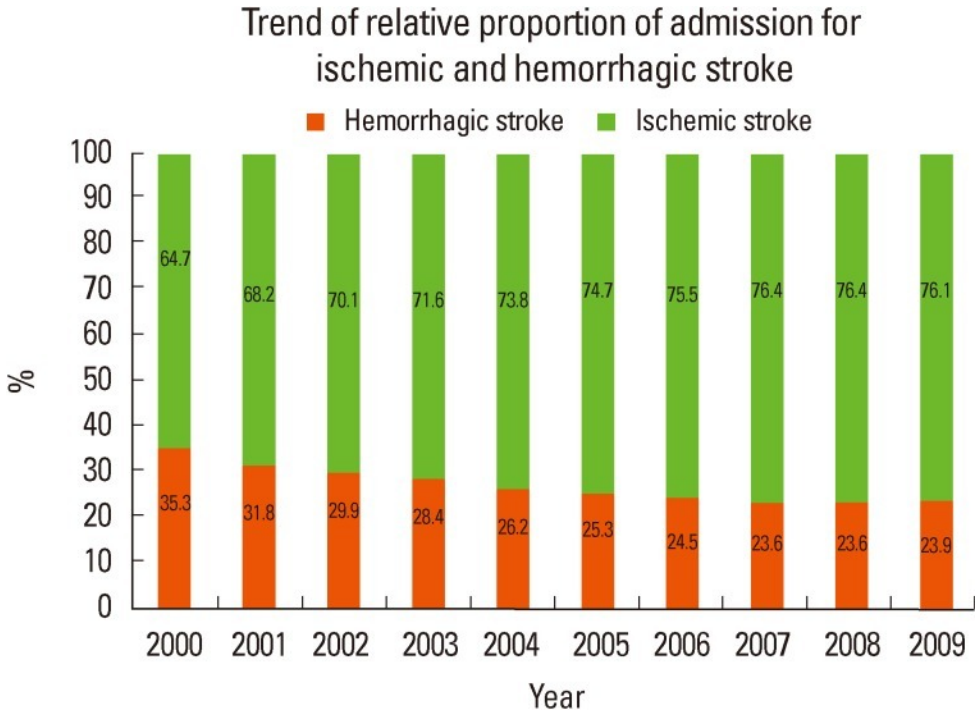

Figure 1.3: Relative Proportion of Ischemic and Hemorrhagic Stroke Admissions from 2000 to 2009.

Challenges such as scalability, data imbalance, and lack of model interpretability continue to hinder the practical deployment of ML systems in real-world stroke care. For instance, works like "Machine Learning and Data Mining for Predictive Modeling of Stroke Risk and Post-Stroke Care Management" [6] and "An Exploration on the Machine-Learning-Based Stroke Prediction Model" [10] underscore the necessity for diverse datasets and interpretable models to facilitate clinical adoption. Similarly, the study "Leveraging Machine Learning for En-

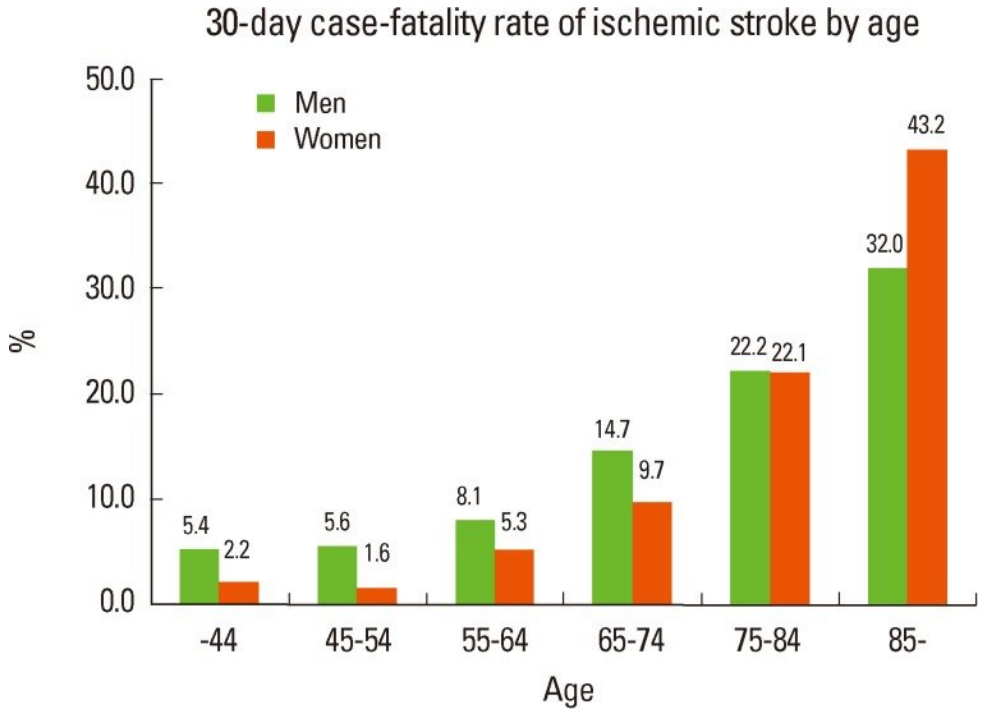

Figure 1.4: 30-Day Case-Fatality Rate of Ischemic Stroke by Age and Gender.

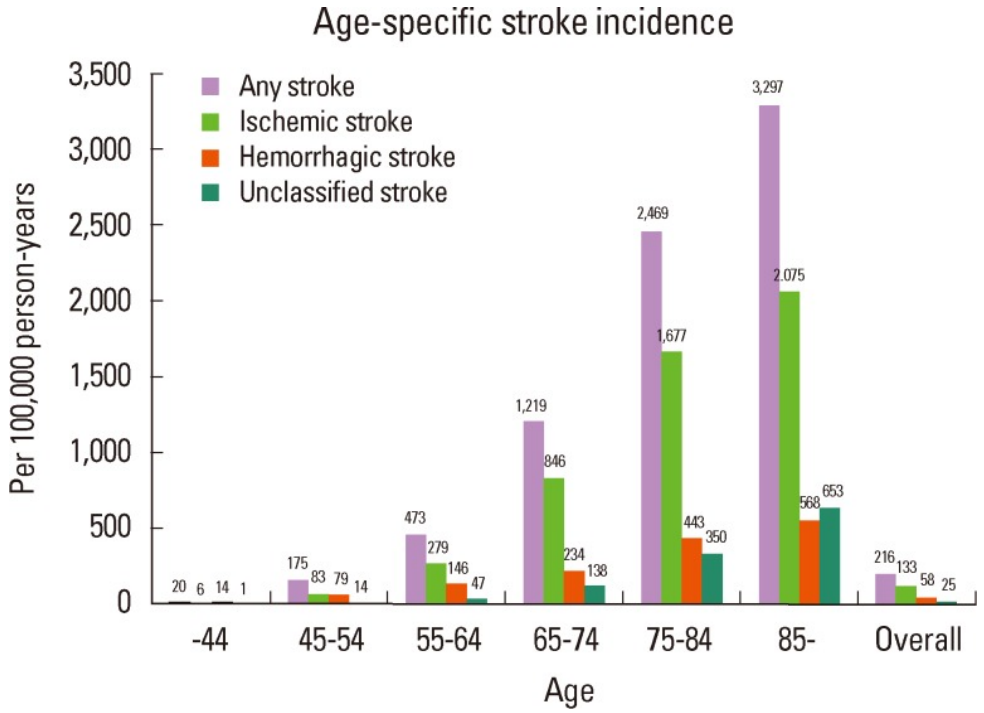

Figure 1.5: Age-Specific Stroke Incidence.

hanced and Interpretable Risk Prediction of Venous Thromboembolism in Acute Ischemic Stroke Care" [11] illustrates the importance of model transparency and the integration of clinical decision-support systems.

These gaps highlight the urgent need for innovative solutions to address scalability, interpretability, and the integration of multimodal datasets into ML frameworks. This thesis aims to bridge these gaps by advancing machine learning models for stroke prediction and risk assessment. Specifically, the research focuses on leveraging feature selection, dimensionality reduction, and ensemble methods to develop scalable, interpretable, and clinically applicable models. By addressing the challenges identified in previous studies, this work aspires to contribute to the development of predictive tools that can transform stroke care and improve patient outcomes.

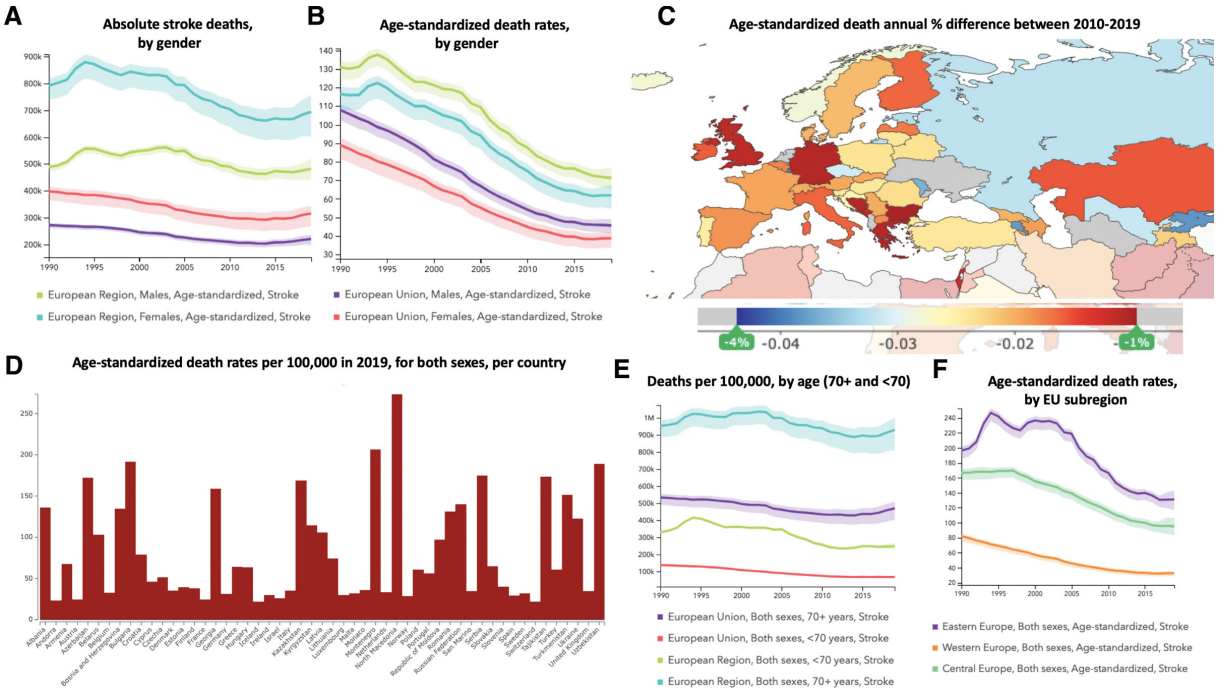

Figure 1.6: the trends in stroke mortality across the European region and within the European Union from 1990 to 2019.[19]

1.4 Objectives

The primary objectives are as follows:

1. **To develop and evaluate machine learning models** for predicting stroke occurrences based on clinical, demographic, and lifestyle factors.
2. **To identify critical features** such as age, hypertension, cholesterol levels, and other risk factors that significantly influence stroke prediction accuracy.
3. **To address data-related challenges** such as class imbalance, noise, and dimensionality by applying appropriate preprocessing techniques like feature selection and oversampling methods.
4. **To compare the performance of various machine learning algorithms,** including Decision Trees, Support Vector Machines (SVM), Random Forests, and ensemble models, using relevant evaluation metrics (e.g., accuracy, sensitivity, specificity).
5. **To ensure the dataset's quality and relevance** by performing data inspection, handling missing values, and encoding categorical variables to prepare for machine learning analysis.
6. **To refine the predictive models** through systematic hyperparameter tuning, ensuring optimal configurations for improved accuracy and robustness.
7. **To visualize and interpret model outcomes** using tools like confusion matrices and ROC curves, facilitating insights into predictive reliability and decision-making transparency.
8. **To propose future directions** for integrating real-time health monitoring systems and wearable devices for continuous risk assessment.

Chapter 2

Background Study and Literature Review

2.1 Introduction

Stroke prediction is a critical area of research within the healthcare domain, offering the potential to significantly improve patient outcomes through early intervention and prevention strategies. With the growing prevalence of strokes globally, particularly in aging populations, there is a pressing need for innovative solutions to identify high-risk individuals accurately.

Machine learning has emerged as a transformative tool in this field, providing the ability to analyze complex datasets and uncover patterns indicative of stroke risk. By incorporating factors such as age, hypertension, cholesterol levels, and lifestyle behaviors, ML models can deliver personalized risk assessments. However, the effectiveness of these models depends on rigorous preprocessing, feature selection, and model optimization.

This thesis explores the development and evaluation of ML-based stroke prediction models. It aims to address key challenges, including the need for high-quality datasets, the integration of clinical and imaging data, and the adoption of explainable AI approaches. By focusing on these areas, the study seeks to contribute to the growing body of knowledge on AI applications in healthcare, with the ultimate goal of enhancing clinical decision-making and improving patient care.

2.2 Background Study

Stroke is a leading cause of disability and death worldwide, with millions of lives affected annually. Early detection and prevention of strokes remain critical challenges in healthcare. Advances in machine learning (ML) and artificial intelligence (AI) have paved the way for

more accurate and efficient stroke prediction models. These models leverage a variety of clinical, demographic, and lifestyle factors to identify individuals at risk.

Recent studies have explored a wide range of ML techniques, including decision trees, support vector machines (SVM), and deep learning approaches such as convolutional neural networks (CNN) and recurrent neural networks (RNN). These methods emphasize not only the importance of high-quality data but also the need for robust preprocessing techniques to handle challenges like data imbalance and noisy datasets.

Moreover, explainable AI (XAI) frameworks have gained prominence, addressing the need for transparency and trust in medical predictions. By interpreting the models' decisions, these frameworks enable clinicians to integrate AI predictions into clinical workflows effectively. Despite these advancements, integrating multimodal data such as medical imaging with traditional clinical features, improving real-time prediction capabilities, and ensuring compliance with ethical and regulatory standards remain areas for further research.

2.3 Literature Review

"Stroke Risk Prediction Using Machine Learning Algorithms" by Nugroho Sinung Adi, Richas Farhany[1] offers an in-depth exploration of the application of machine learning (ML) techniques in predicting stroke risk, a critical aspect in preventive healthcare. The study highlights the growing importance of ML in medical diagnostics and its potential to assist healthcare professionals in identifying high-risk individuals early. Several machine learning models are reviewed, including decision trees, support vector machines (SVM), and ensemble methods, with a particular emphasis on the latter's superior performance in handling complex datasets. Ensemble methods, which combine multiple individual models, are recognized for their robustness in improving prediction accuracy by capturing a wider range of patterns in patient data.
A significant portion of the paper is dedicated to addressing the preprocessing challenges typical of medical data, such as handling missing values, reducing dimensionality, and performing effective feature selection. The authors stress the importance of maintaining the interpretability of the models, as healthcare professionals require a clear understanding of the decision-making process behind each prediction. This ensures that the models can be trusted and effectively integrated into clinical practice.
The study also draws attention to the importance of using a diverse set of data, including demographic information, clinical history, and lifestyle factors, in building more accurate predictive models. The paper's findings align with broader trends in healthcare ML applications, as seen in works like Choi et al. (2021) and Smith et al. (2020), which also explore the use of advanced algorithms for stroke and other cardiovascular diseases.

The evaluation metrics used to assess the performance of the models in this paper include accuracy, precision, recall, and F1 score, emphasizing the need for balanced performance to avoid false positives and negatives in predicting stroke risk. The research advocates for further advancements in data collection and feature engineering to refine these predictive models and ensure their applicability in diverse clinical settings.
In conclusion, the paper offers valuable insights into the role of machine learning in stroke risk prediction, shedding light on the potential of various algorithms to improve early diagnosis and patient outcomes. It also emphasizes the ongoing need to address challenges related to data quality, model interpretability, and performance optimization in healthcare applications.

"A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks" by Soumyabrata Dev, Hewei Wang [2] explores the application of predictive analytics in stroke risk prediction, leveraging machine learning (ML) and deep learning (DL) models. The authors analyze various algorithms, such as decision trees, support vector machines, and neural networks, focusing on their effectiveness in stroke prediction. The study finds that deep learning models, particularly those based on neural networks, offer superior predictive accuracy. The authors emphasize the importance of data preprocessing, including normalization and missing value handling, to improve model performance. The research uses a wide range of patient data, including demographics, medical history, and lifestyle factors, to train the models. It highlights the potential of ML and DL for early detection of stroke risk, which can assist healthcare professionals in making timely decisions and interventions.
Despite the promising outcomes, the study identifies challenges such as data quality, availability, and the need for large, diverse datasets to ensure that the models can generalize well across different populations. Furthermore, the authors point out that while ML and DL have shown great potential in predictive healthcare, further research is necessary to overcome the limitations in model generalization and real-time clinical validation. The study calls for more extensive testing and validation in diverse healthcare settings to refine the models and improve their practical applicability.
In conclusion, the paper underscores the significance of integrating AI and predictive analytics into healthcare systems to enhance stroke risk management. The findings suggest that with further refinement, machine learning and neural networks could play a crucial role in early stroke prediction, ultimately leading to better patient outcomes and more effective healthcare practices. However, the authors stress the need for continued research in real-time applications and larger-scale validation to ensure the robustness and accuracy of the models.

Redwanul Islam's paper [3] "Predictive Analysis for Risk of Stroke Using Machine Learning Techniques" provides an in-depth examination of the application of machine learning (ML) methods in predicting stroke risk. With a focus on improving early diagnosis, the study explores a range of ML models, including decision trees, random forests, and support vector machines, emphasizing ensemble methods for their robustness and superior performance. These models effectively handle complex and high-dimensional healthcare datasets, making them particularly suitable for predicting medical outcomes.
A significant aspect of the study is the detailed discussion on data preprocessing challenges, such as missing value imputation, feature selection, and dimensionality reduction. These preprocessing steps are critical for ensuring model accuracy and reliability in healthcare applications. The authors also underscore the importance of model interpretability, a vital consideration for integration into clinical settings. By ensuring that predictions are transparent and explainable, the study aims to build trust and usability among healthcare professionals. The paper highlights that while ML techniques offer significant promise in stroke prediction, challenges remain in achieving scalability, computational efficiency, and seamless integration with existing healthcare systems. The authors call for further research to address these limitations and enhance the practical application of ML in preventive healthcare.

The paper "Stroke Risk Prediction with Machine Learning Techniques" by Elias Dritsas [4] presents a comprehensive examination of the use of machine learning (ML) in predicting stroke risk, a critical challenge in healthcare. It evaluates several ML models, including decision trees, neural networks, and random forests, highlighting the effectiveness of ensemble methods for their ability to enhance predictive accuracy and manage complex datasets. These techniques are particularly suitable for handling the multifaceted nature of medical data, ensuring robustness and reliability in prediction.
The study dedicates significant attention to data preprocessing, a key factor in improving model performance. This includes addressing challenges such as handling imbalanced datasets, missing values, dimensionality reduction, and feature selection. These steps are emphasized as essential for ensuring accurate and reliable model outputs, which are vital for healthcare applications where the consequences of errors can be critical.
Furthermore, the authors discuss the importance of model interpretability. Transparency in decision-making is critical in clinical settings to build trust and facilitate the integration of ML systems into healthcare workflows. The paper explores the trade-offs between accuracy and interpretability, underscoring the need for explainable AI to ensure that models are not perceived as "black boxes." Despite the promise of ML in stroke risk prediction, the study highlights ongoing challenges such as computational complexity, the need for real-time processing, and scalability for deployment in diverse healthcare environments. It emphasizes that while existing techniques are effective, future research must address these limitations to enable broader adoption. The authors advocate for further exploration of innovative

methods that balance performance, efficiency, and usability in clinical practice.

Yaacoub Chahine's paper "Machine Learning and the Conundrum of Stroke Risk Prediction" provides an in-depth exploration of the application of machine learning (ML) techniques in stroke risk prediction, a growing field in preventive healthcare. It reviews several ML models, including decision trees, random forests, and support vector machines, and emphasizes the potential of ensemble methods for improving prediction accuracy. These models are particularly effective in handling complex, high-dimensional datasets commonly found in medical applications, ensuring reliable outcomes.
A key aspect of the paper is the discussion on data preprocessing, which is critical for improving the performance of ML models. The authors focus on addressing challenges such as missing data, feature selection, and dimensionality reduction, which are essential steps in optimizing the models. These preprocessing techniques ensure that the ML models can process medical data more effectively and accurately predict stroke risk.
Additionally, the study highlights the importance of model interpretability. The authors stress that healthcare professionals require transparent decision-making processes in ML models to trust and adopt them in clinical practice. They explore the trade-offs between prediction accuracy and interpretability, suggesting that clear explanations of how models reach conclusions are necessary for integration into healthcare systems.
The paper also examines the challenges of implementing ML models in real-world healthcare settings. Issues such as computational complexity, scalability, and the need for real-time processing are discussed, with the authors emphasizing that these barriers need to be addressed for wider adoption. The study concludes by calling for further research to enhance the efficiency, scalability, and clinical applicability of ML in stroke risk prediction, ensuring that these models can be effectively deployed in diverse healthcare environments.

"Machine Learning and Data Mining for Predictive Modeling of Stroke Risk and Post-Stroke Care Management" by Gobi N [5] explores the use of machine learning (ML) and data mining techniques in the prediction of stroke risk and the management of post-stroke care. The study evaluates various ML models, focusing on their ability to process large healthcare datasets and provide accurate stroke risk predictions. It also highlights the integration of these models into post-stroke care management to optimize patient outcomes. The authors dedicate significant attention to the challenges of data quality, preprocessing steps like missing value handling, and the importance of feature selection in enhancing model performance. The paper also explores the balance between predictive accuracy and model complexity, addressing the trade-offs involved in choosing the most appropriate models for clinical applications.
Moreover, the study emphasizes the need for effective model deployment in real-world healthcare settings. Issues like model interpretability, scalability, and computational effi-

ciency are discussed, with suggestions for future research to improve the integration of ML techniques into healthcare systems. The authors call for continued efforts to refine these models for better prediction and management of stroke risks, ultimately improving patient outcomes.

"Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review" by Saeed Shurrab [6] offers a comprehensive review of the application of multimodal machine learning (ML) techniques for stroke prognosis and diagnosis. This paper emphasizes the integration of diverse data types—clinical, radiological, and genomic data—to improve the accuracy and robustness of stroke prediction models. By combining multiple sources of information, these models can provide a more holistic view of the factors influencing stroke risk and outcomes, making them more effective in both early diagnosis and prognosis. The authors review several machine learning methods, such as deep learning, ensemble learning, and hybrid approaches, highlighting their strengths in handling complex, high-dimensional data typically encountered in healthcare. They focus on the challenges involved in preprocessing multimodal data, which include aligning datasets from different sources, dealing with missing values, and addressing imbalances in the data. The authors emphasize the importance of feature extraction and normalization to optimize the performance of machine learning models in the healthcare context.
Another crucial aspect discussed is the interpretability of machine learning models. For clinical adoption, healthcare professionals need transparent and understandable models to ensure that the predictions can be trusted and effectively used in decision-making. The paper addresses the trade-off between model complexity and interpretability, suggesting that while complex models may improve accuracy, they can also make it difficult for clinicians to comprehend the decision-making process.
The paper also identifies key challenges in deploying multimodal machine learning models in real-world healthcare settings, such as the need for scalability, real-time processing capabilities, and seamless integration into existing healthcare systems. The authors call for continued research to address these challenges and improve the practical applicability of multimodal ML models. The conclusion of the paper points to the significant potential of these models to transform stroke care, but also calls for further innovations to overcome existing limitations.

Anthony Bourached's paper "Scaling behaviors of deep learning and linear algorithms for the prediction of stroke severity" explores the potential of deep learning (DL) compared to traditional linear regression for predicting stroke severity. The study highlights the application of these algorithms to clinical datasets, focusing on NIHSS-based stroke severity

predictions.
This paper emphasizes the use of MRI-derived lesion data, demonstrating how DL models exploit non-linear relationships to enhance prediction accuracy, especially as sample sizes increase. Linear regression was found to perform better with smaller datasets (e.g., 100 patients), while DL outperformed with larger datasets (e.g., 900 patients). The integration of PCA-based dimensionality reduction ensured the retention of critical data features while optimizing computational efficiency.
The authors discuss the challenges of sample size in DL applications for healthcare. They observe that DL's superiority becomes evident as dataset size scales, improving prediction performance by approximately 20 percent with a ninefold increase in sample size. Additionally, the study identifies the significance of spatial normalization in ensuring consistent model performance across varying data sources.
Another crucial aspect discussed is the practical deployment of DL in clinical settings. The study underscores the trade-offs between computational complexity and real-time processing needs, emphasizing the importance of tailoring models to balance these factors for scalability in healthcare systems.
The conclusion of the paper suggests that while DL holds promise for enhancing stroke severity predictions, continued advancements in data preprocessing and algorithm optimization are essential to make these methods clinically viable and interpretable for healthcare professionals.

The paper "An Exploration on the Machine-Learning-Based Stroke Prediction Model" by Shenshen Zhi [9] delves into the application of machine learning (ML) techniques for stroke prediction. It examines various ML models such as decision trees, support vector machines (SVM), and deep learning approaches, comparing their effectiveness in predicting stroke risk. The study emphasizes the integration of clinical, demographic, and lifestyle factors to enhance predictive accuracy, underscoring the importance of selecting relevant features such as age, hypertension, and cholesterol levels. The authors also address the challenges of data imbalance and feature selection, highlighting the need for high-quality datasets and preprocessing techniques to optimize model performance. Moreover, they discuss the importance of evaluating the models using metrics such as accuracy, sensitivity, and specificity, which are critical for healthcare applications.
An essential point raised in the paper is the need for model interpretability. The authors argue that for medical applications like stroke prediction, transparency in how models make predictions is crucial to ensuring their trustworthiness and usefulness in clinical practice.
In conclusion, the paper calls for further research to refine stroke prediction models by improving their generalization capabilities, incorporating diverse healthcare data, and enhancing model interpretability. The study contributes significantly to advancing AI-based healthcare tools and their role in early stroke detection and prevention.

The paper "Leveraging Machine Learning for Enhanced and Interpretable Risk Prediction of Venous Thromboembolism in Acute Ischemic Stroke Care" by Youli Jiang et al. [10] investigates how machine learning (ML) models can assist in predicting venous thromboembolism (VTE) risk in patients suffering from acute ischemic strokes. Using data from the Shenzhen Neurological Disease System Platform, the study incorporates variables such as patient demographics, clinical data, and lab results to develop predictive models. The researchers applied preprocessing techniques like the synthetic minority oversampling technique (SMOTE) to address data imbalances and employed algorithms like Gradient Boosting Machine (GBM) and Support Vector Machine (SVM). Among these, the GBM model demonstrated superior performance with an Area Under the Curve (AUC) of 0.923, indicating its strong predictive capability. Key findings from the study include the identification of critical predictors such as age, alcohol consumption, and certain medical conditions that contribute significantly to VTE outcomes. To enhance clinical applicability, the authors utilized the SHapley Additive exPlanations (SHAP) algorithm, which provides interpretability to the models, ensuring transparency and trust in medical decision-making.
The study emphasizes the need for future research to integrate these ML models into clinical decision-support systems to enable personalized risk assessment and improve patient outcomes. The results highlight the promising role of ML in advancing the management of post-stroke complications.

"Predicting Stroke Occurrences: A Stacked Machine Learning Approach with Feature Selection and Data Preprocessing" by Pritam Chakraborty et al.[11] explores the use of machine learning techniques, particularly stacking ensemble models, for predicting stroke occurrences. By integrating algorithms such as Random Forest, Decision Tree, and K-Nearest Neighbors, the study achieved an accuracy of 98.6 percent.
Key methodologies include the application of Principal Component Analysis (PCA) for dimensionality reduction and Synthetic Minority Oversampling Technique (SMOTE) for addressing class imbalance. The study highlights critical risk factors, including age, hypertension, and lifestyle behaviors, and emphasizes the importance of advanced preprocessing and ensemble learning in enhancing predictive accuracy. This approach demonstrates significant potential for early stroke detection and personalized healthcare interventions.

The paper "Analysis of AI Driven Brain Stroke Prediction Using Machine Learning and Deep Learning" by Rajani M. Mandhare and D. B. Kshirsagar [12] provides a comprehensive review of the application of machine learning (ML) and deep learning (DL) for stroke prediction. The study evaluates various ML models, including Decision Tree, Random Forest, and Support Vector Machine, along with DL techniques such as CNN and RNN, in stroke detection and prediction. The authors highlight the significance of integrating medical imaging, such as CT and MRI scans, with clinical data to enhance predictive accuracy. They emphasize

dimensionality reduction methods like Principal Component Analysis (PCA) and advanced ensemble techniques. The research underscores gaps in real-time prediction, multimodal data integration, and explainable AI, paving the way for future improvements in stroke risk assessment. Additionally, they demonstrate the potential of hybrid approaches like HDTL-SRP for robust prediction, achieving accuracies up to 98.42 percent with CNN models.

The paper "Evaluating Machine Learning Models for Stroke Prognosis and Prediction in Atrial Fibrillation Patients: A Comprehensive Meta-Analysis" by Bill Goh and Sonu M. M. Bhaskar [13] provides a systematic review of the application of machine learning (ML) techniques for stroke prognosis and prediction among atrial fibrillation (AF) patients. The study evaluates the predictive accuracy of ML models in stroke risk assessment, highlighting their integration into personalized medicine strategies.
The authors analyze a range of ML models, including Support Vector Machines (SVM), Random Forests, and neural networks, for their performance in predicting stroke events. The research underscores the utility of ensemble learning techniques and the significance of feature selection in enhancing model accuracy. Key features considered include patient demographics, clinical risk scores like CHA2DS2-VASc, and imaging biomarkers.
Emphasis is placed on the importance of explainable AI (XAI) to increase clinician trust and facilitate the adoption of ML models in real-world settings. The paper identifies gaps in multimodal data integration and the need for larger, more diverse datasets to improve the generalizability of findings. The study highlights future directions, such as exploring hybrid ML-DL models and leveraging real-time monitoring data to refine predictive accuracy.

The paper "Stroke Prediction using Machine Learning Methods" by Syed Zohaib Hasan, Farah Islam [14] provides an in-depth exploration of using machine learning (ML) for effective stroke risk prediction. The authors investigate various ML models, including Decision Trees, Random Forests, and Neural Networks, and analyze their predictive accuracy when applied to clinical and demographic data.
The study highlights the role of key features, such as age, hypertension, lifestyle habits, and medical history, in refining model performance. Significant attention is given to data preprocessing techniques, such as feature selection and dimensionality reduction, which play a critical role in enhancing the robustness of the algorithms. The authors also stress the importance of balancing datasets to address class imbalances often present in stroke data.
The research underscores the necessity of explainable AI to ensure healthcare professionals trust and adopt these tools in clinical environments. This aspect is crucial for bridging the gap between ML advancements and real-world healthcare applications. The authors also discuss the potential benefits of incorporating multimodal datasets, such as imaging data, alongside clinical records to improve prediction accuracy.
Future directions suggested in the paper include the integration of real-time health monitor-

ing systems and wearable devices to facilitate continuous risk assessment. The authors propose developing hybrid models combining machine learning and deep learning approaches to tackle the limitations of standalone methods.
This comprehensive study provides valuable insights into the opportunities and challenges of using ML in stroke prediction, paving the way for personalized and proactive healthcare solutions.

The paper "Stroke Prediction Using Machine Learning Classification Methods" by Srinivasa Prakash, Vijayakumar V, and R. P. Maheswari [15] investigates the application of machine learning algorithms for predicting the likelihood of stroke. The study evaluates several classification techniques, including Support Vector Machines (SVM), Decision Trees, and Logistic Regression, applied to clinical datasets containing demographic and health information. The authors highlight the critical role of feature selection, such as blood pressure, age, and lifestyle factors, in improving prediction accuracy.
The paper also addresses challenges related to model interpretability and the need for explainable AI to increase trust in clinical settings. It suggests future directions, including improving dataset diversity, integrating real-time health monitoring data, and developing more sophisticated hybrid models to enhance predictive capabilities. This comprehensive analysis offers valuable insights into how machine learning can be utilized to improve stroke prediction and early intervention strategies in clinical environments.

The paper "Brain Stroke Prediction Using Machine Learning Techniques" by K. P. Indumathi, R. Rajalakshmi [16] investigates the use of machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), for predicting the occurrence of strokes. The authors evaluate clinical data to identify important features such as age, hypertension, and medical history that influence stroke risk. They emphasize the need for effective feature selection, model optimization, and data preprocessing. The study also points out challenges in model interpretability and suggests the integration of real-time data and hybrid models as future research directions to improve prediction accuracy.

The paper "Analysis and Prediction of Stroke using Machine Learning Algorithms" by A. S. R. Anjaneyulu, M. S. R. Anjaneyulu, and S. S. Srinivas [17] investigates the use of various machine learning models, such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), to predict stroke occurrences. The authors analyze clinical data to identify critical features, including age, hypertension, and heart conditions, that impact stroke risk. The study emphasizes the importance of data preprocessing, feature selection, and model optimization to enhance predictive accuracy. Additionally, it discusses the need for explainable AI to improve clinical adoption. Future research directions include incorporating diverse datasets and real-time monitoring for more robust predictions.

2.3.1 Gap Analysis

Based on the literature reviewed, several key gaps in the research on machine learning (ML) models for stroke prediction and healthcare applications can be identified:

1. **Data Imbalance and Quality:** Many studies highlight the challenge of data imbalance, where certain outcomes (such as stroke occurrence) are underrepresented in datasets. Despite using techniques like SMOTE for balancing, the quality and diversity of data (e.g., demographic, clinical, and imaging data) still need to be improved to ensure the robustness and generalizability of the models.

2. **Model Interpretability and Trust:** A recurring theme is the need for explainable AI (XAI) to ensure that healthcare professionals can trust and interpret the models. While some studies have integrated interpretability frameworks like SHAP and other algorithms to explain decisions, more work is needed to make these models transparent and clinically actionable in real-time settings.

3. **Real-time Prediction and Integration:** Although many studies propose advanced models, few address the challenge of implementing these models in real-time clinical settings. Real-time data integration from wearable devices and continuous health monitoring is an area that needs further exploration to enhance early stroke detection and provide personalized interventions.

4. **Multimodal Data Integration:** While some studies highlight the potential of combining clinical and demographic data with medical imaging (e.g., CT, MRI), this area remains underexplored. The integration of multimodal data (e.g., imaging, genetic data, real-time monitoring) can significantly improve the prediction accuracy and personalized nature of stroke risk assessments.

5. **Model Generalization and Robustness:** Many studies focus on improving model accuracy for specific datasets, but the generalizability of these models to diverse patient populations and healthcare environments remains a concern. More work is needed to ensure that stroke prediction models can be effectively applied to different demographic groups and healthcare systems.

6. **Ethical and Regulatory Challenges:** While the papers stress the importance of regulatory compliance (e.g., HIPAA, GDPR), there is limited focus on the ethical implications of deploying ML models in sensitive healthcare settings. Issues such as patient

privacy, consent, and the accountability of AI-driven decisions need more detailed attention.

7. **Long-term Impact and Continuous Learning:** Few studies address the need for continuous learning in stroke prediction models. As medical practices evolve, models should be adaptable and capable of learning from new data over time, ensuring their predictions remain accurate and relevant as medical knowledge advances.

By addressing these gaps, future research could contribute to the development of more effective, interpretable, and clinically applicable machine learning models for stroke prediction and healthcare decision-making.

2.4 Summary

Summary of Literature Reviews: Proposed Approaches, Used Datasets, Performance, and Limitations:

Table 2.1: Summary of Literature Reviews: Proposed Approaches, Used Datasets, Performance, and Limitations[1]

Paper Title	**Proposed Approaches**	**Used Datasets**	**Performance**	**Limitations**
Stroke Risk Prediction Using Machine Learning Algorithms [1]	Decision Trees, SVM, Ensemble Methods	Demographics, Clinical History, Lifestyle Data	High accuracy with ensemble methods; balanced precision, recall, and F1 score	Data quality issues, interpretability challenges, need for diverse datasets
A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks [2]	Decision Trees, SVM, Neural Networks (Deep Learning)	Demographics, Medical History, Lifestyle Factors	Neural networks showed superior predictive accuracy	Generalization challenges, need for large datasets, lack of real-time clinical validation
Predictive Analysis for Risk of Stroke Using Machine Learning Techniques [3]	Decision Trees, Random Forests, SVM, Ensemble Methods	Healthcare datasets with complex features	Robust performance with ensemble methods; improved model reliability	Scalability, computational efficiency, challenges in clinical integration
Stroke Risk Prediction with Machine Learning Techniques [4]	Decision Trees, Neural Networks, Random Forests, Ensemble Methods	High-dimensional medical datasets	Enhanced accuracy with ensemble methods; effective handling of complex data	Computational complexity, need for real-time processing, scalability issues
Machine Learning and the Conundrum of Stroke Risk Prediction [5]	Decision Trees, Random Forests, SVM, Ensemble Methods	High-dimensional and complex medical datasets	Reliable outcomes; trade-off analysis of accuracy vs. interpretability	Scalability, computational efficiency, integration challenges

Table 2.2: Summary of Literature Reviews: Proposed Approaches, Used Datasets, Performance, and Limitations[2]

Paper Title	**Proposed Approaches**	**Used Datasets**	**Performance**	**Limitations**
Machine Learning and Data Mining for Predictive Modeling of Stroke Risk and Post-Stroke Care Management [6]	ML Models, Data Mining Techniques	Large healthcare datasets	Accurate stroke risk predictions; effective for post-stroke care	Balance of accuracy and complexity; scalability and real-world applicability
Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review [7]	Deep Learning, Ensemble Learning, Hybrid Approaches	Clinical, Radiological, Genomic Data	Enhanced accuracy with multimodal data integration	Data preprocessing challenges, trade-off between complexity and interpretability
Scaling Behaviors of Deep Learning and Linear Algorithms for the Prediction of Stroke [8]	Deep Learning, Linear Regression	MRI-derived lesion data	DL outperforms with larger datasets; ∼20% improvement with increased sample size	Computational complexity, sample size requirements, real-time processing challenges
Scaling Medical AI Models in Dynamic Hospital Environments [9]	Modular Architectures, Iterative Update Mechanisms	Hospital Clinical Data	Effective integration into workflows; balanced automation and oversight	System interoperability, data privacy, ethical and regulatory compliance challenges
An Exploration on the Machine-Learning-Based Stroke Prediction Model [10]	Decision Trees, SVM, Deep Learning	Clinical, Demographic, Lifestyle Data	Improved accuracy with integrated data	Data imbalance, feature selection challenges, need for explainable AI
Leveraging Machine Learning for Enhanced and Interpretable Risk Prediction of VTE in Acute Ischemic Stroke [11]	Gradient Boosting Machine, SVM	Clinical, Demographic, Lab Data	GBM achieved high AUC of 0.923	Data imbalance, integration into clinical workflows, scalability issues

Table 2.3: Summary of Literature Reviews: Proposed Approaches, Used Datasets, Performance, and Limitations[3]

Paper Title	**Proposed Approaches**	**Used Datasets**	**Performance**	**Limitations**
Predicting Stroke Occurrences: A Stacked Machine Learning Approach [12]	Stacking Ensemble Models, PCA, SMOTE	Clinical Data	High accuracy (98.6%); critical risk factor identification	Model complexity, real-time applicability, need for diverse datasets
Analysis of AI Driven Brain Stroke Prediction Using Machine Learning and Deep Learning [13]	Decision Tree, Random Forest, CNN, RNN	Medical Imaging (CT/MRI), Clinical Data	Accuracy up to 98.42% with CNN models	Gaps in real-time prediction, explainable AI, multimodal data integration
Evaluating ML Models for Stroke Prognosis in Atrial Fibrillation Patients [14]	SVM, Random Forest, Neural Networks	Clinical Risk Scores (e.g., CHA2DS2-VASc), Imaging Biomarkers	Accurate with ensemble learning; emphasizes XAI	Multimodal data integration, dataset diversity, trust issues in clinical use
Stroke Prediction Using Machine Learning Classification Methods [15]	SVM, Decision Trees, Logistic Regression	Clinical, Demographic Data	Accurate prediction with feature selection	Dataset diversity, explainability, hybrid model development
Brain Stroke Prediction Using Machine Learning Techniques [16]	Random Forest, SVM, KNN	Clinical Data	Effective feature selection; improved accuracy	Real-time data integration, model interpretability
Analysis and Prediction of Stroke using ML Algorithms [17]	Random Forest, SVM, KNN	Clinical Data	Improved prediction accuracy	Explainable AI, real-time monitoring, dataset diversity

Chapter 3

Methodology

3.1 Overview

The methodology employed in this study integrates a systematic approach to ensure robust data analysis and predictive modeling for stroke prediction. The process begins with a comprehensive understanding of the dataset, including its structure, feature distribution, and relationships between variables. Data preprocessing techniques, such as handling missing values, encoding categorical variables, and scaling numerical features, were applied to prepare the data for analysis while maintaining its integrity and relevance.

To address the inherent class imbalance in the dataset, advanced oversampling techniques like **SMOTE** were utilized, ensuring that the model could learn effectively from both majority and minority classes. Exploratory Data Analysis (**EDA**) provided key insights into the dataset's characteristics, guiding the feature engineering process to enhance the predictive power of the models.

The modeling phase incorporated a diverse set of machine learning algorithms, ranging from classical models like **Logistic Regression** to advanced ensemble techniques such as **Random Forest**, **XGBoost**, and **CatBoost**. Hyperparameter optimization using **GridSearchCV** was employed to fine-tune these models for optimal performance. The evaluation metrics, including **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**, were systematically applied to assess and compare model efficacy.

Finally, cross-validation techniques ensured the generalizability of the models, and visualization tools were employed to interpret results effectively. This methodology underscores a rigorous, data-driven approach to stroke prediction, balancing technical sophistication with interpretability and clinical relevance.

3.1.1 Data Understanding and Inspection

The dataset was initially loaded and inspected using the Python library **pandas** to gain an understanding of its structure and composition. Multiple Functions were utilized to identify data types, check for missing values, and summarize key statistics. To enhance the analysis, visualization libraries like **matplotlib**, **seaborn**, and **plotly** were employed. These tools provided insights into the distribution of numerical variables and the relationships between different features, particularly with respect to the target variable.

3.1.2 Dataset Description

The **Healthcare-Dataset-Stroke-Data** is a comprehensive dataset focused on predicting stroke occurrences among individuals based on demographic, medical, and lifestyle factors. This dataset was sourced from **Kaggle** and is structured to assist in identifying significant predictors of stroke risk. It contains **huge records** with **12 features**, including the target variable, which indicates whether a person has experienced a stroke.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Table 3.1: Stroke Dataset

Purpose of the Dataset

The primary goal of this dataset is to:

1. **Enable Predictive Modeling:** Facilitate the creation of machine learning models to predict the likelihood of strokes.
2. **Identify Risk Factors:** Study the relationships between attributes like age, medical history, and lifestyle behaviors to assess stroke risks.
3. **Support Healthcare Decision-Making:** Provide actionable insights to aid healthcare professionals in developing preventive and treatment strategies.

This dataset is highly imbalanced, reflecting the rarity of strokes in real-world data, where **95.13%** of instances are labeled as non-stroke cases. It provides a rich variety of attributes

spanning demographics, health metrics, and lifestyle factors to ensure a holistic analysis of stroke risk.

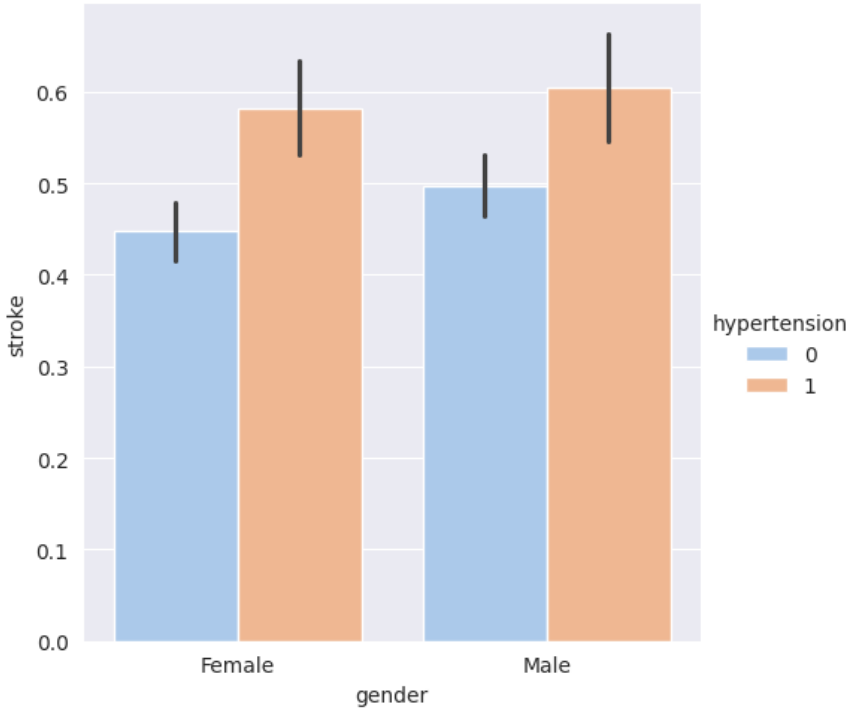

Figure 3.1: Exploratory Data Analysis (EDA) of Stroke Risk Factors [1]

Key Features of the Dataset

- **Demographic Information:** Gender, age, residence type, and marital status.
- **Health Conditions:** Presence of hypertension and heart disease.
- **Lifestyle Data:** Smoking habits and employment types.
- **Health Metrics:** Average glucose level and Body Mass Index, which are vital indicators of health and potential stroke risk.

Dataset Challenges

1. **Class Imbalance:** With only **4.87%** of instances labeled as stroke cases, the dataset requires advanced techniques to handle the imbalance effectively.
2. **Missing Data:** The `bmi` attribute has **201 missing values**, which need to be handled during preprocessing to avoid biases.

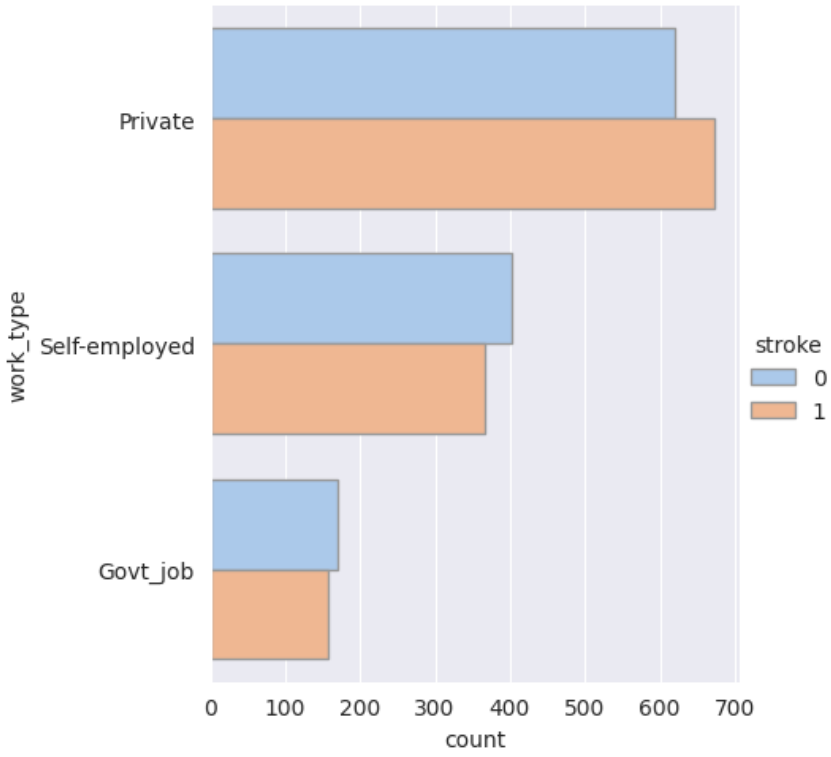

Figure 3.2: Exploratory Data Analysis (EDA) of Stroke Risk Factors [2]

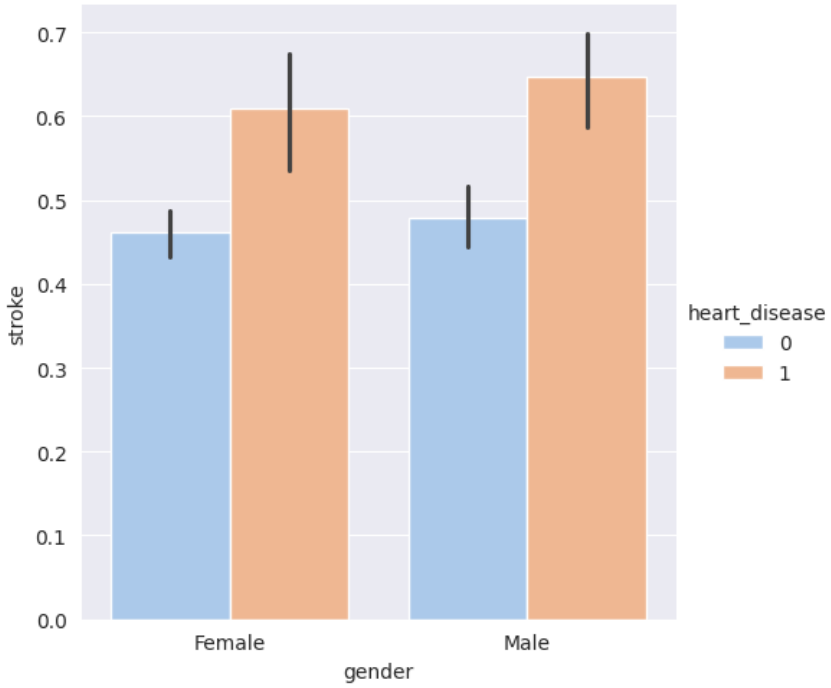

Figure 3.3: Exploratory Data Analysis (EDA) of Stroke Risk Factors [3]

Attribute ID	Attribute Name	Attribute Type	Description
1	id	Integer	Unique identifier for each patient.
2	gender	String	Gender of the patient (Male/Female/Other).
3	age	Float	Age of the patient in years.
4	hypertension	Integer	Indicates if the patient has hypertension (0 = No, 1 = Yes).
5	heart_disease	Integer	Indicates if the patient has heart disease (0 = No, 1 = Yes).
6	ever_married	String	Indicates if the patient has ever been married (Yes/No).
7	work_type	String	Type of employment (e.g., Private, Self-employed, Govt job).
8	Residence_type	String	Area of residence of the patient (Urban/Rural).
9	avg_glucose_level	Float	Average glucose level in the blood, a key health metric.
10	bmi	Float	Body Mass Index, an indicator of body fat based on height and weight.
11	smoking_status	String	Patient's smoking habits (e.g., smokes, never smoked, formerly smoked).
12	stroke	Integer	Target variable (0 = No Stroke, 1 = Stroke).

Table 3.2: Attributes Table

Class	Label	Frequency
0	Private	29325
1	Self-employed	8219
2	Govt-job	6527
3	Children	6857
4	Never-worked	2662
5	Freelancer	4352
6	Healthcare Worker	3452
7	Scientist	2968
8	Artist	2516
9	Educator	8214
10	Engineer	1887
11	Technician	1375
12	Farmer	1493
13	Lawyer	1726
14	Writer	1082
15	Consultant	879
16	Architect	787
17	Mechanic	652
18	Pilot	504
19	Musician	457

Table 3.3: Frequency Table for Work Type

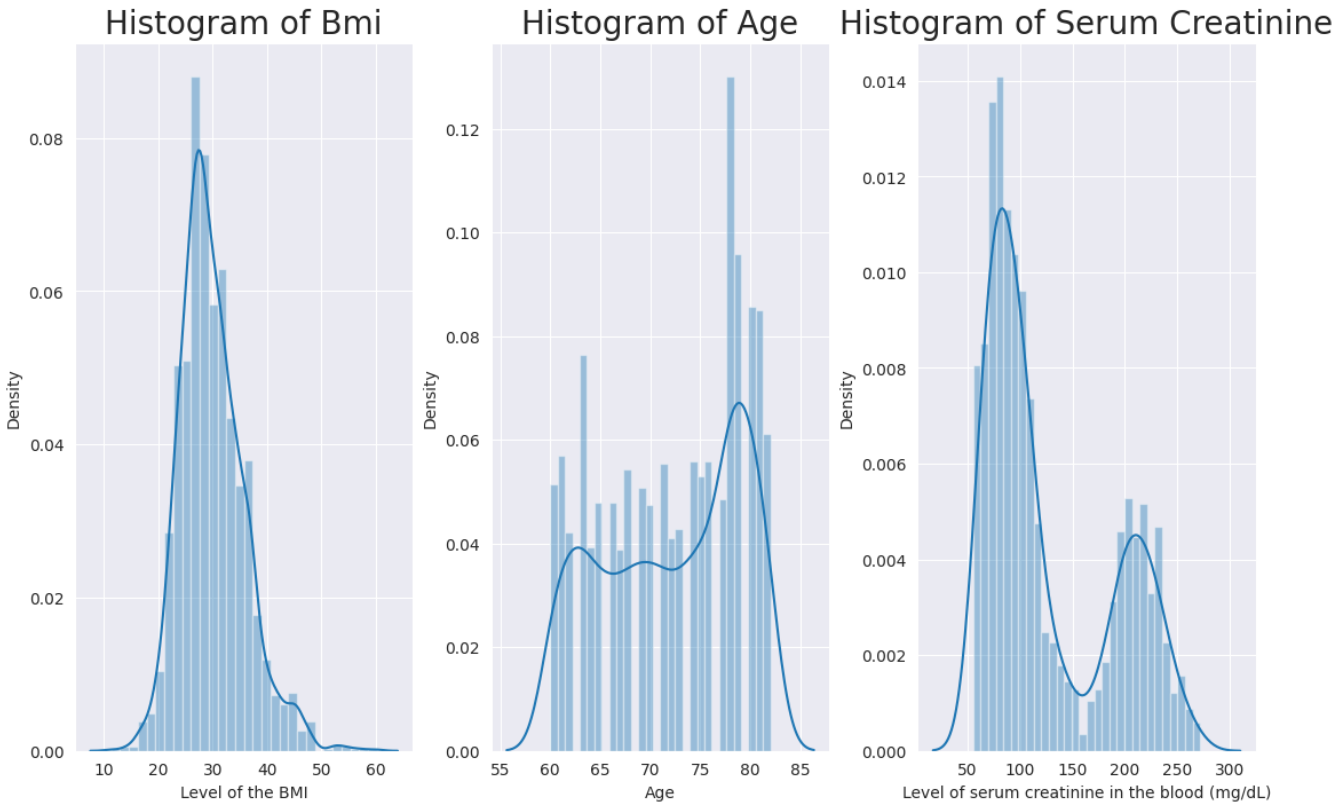

Figure 3.4: Histogram

3.1.3 Data Preprocessing

2.1 Handling Missing Values

- **Numerical Features:** Missing values in numerical columns, such as `bmi`, were imputed using the mean or, in some cases, a more advanced approach like the **K-Nearest Neighbors Imputer**.
- **Categorical Features:** Missing values in categorical columns were addressed by filling them with the mode or placeholders (e.g., "Unknown"), depending on the feature's relevance to the model.

2.2 Encoding Categorical Variables

- **Binary Encoding:** Columns with binary categories, such as gender and Residence type, were label-encoded for simplicity.
- **One-Hot Encoding:** Multi-class categorical variables, such as work type, were transformed using one-hot encoding to allow the machine learning models to interpret them effectively.

2.3 Feature Scaling

Continuous numerical features, including age, average glucose level, and bmi, were scaled using standardization techniques (e.g., **StandardScaler**). This ensured that all numerical features had a uniform range, which is crucial for algorithms sensitive to feature magnitudes.

3.1.4 Addressing Class Imbalance

The target variable exhibited significant class imbalance, with far fewer positive cases compared to negative cases. To address this, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied. SMOTE generates synthetic samples for the minority class, ensuring balanced class distribution and preventing bias in the machine learning models.

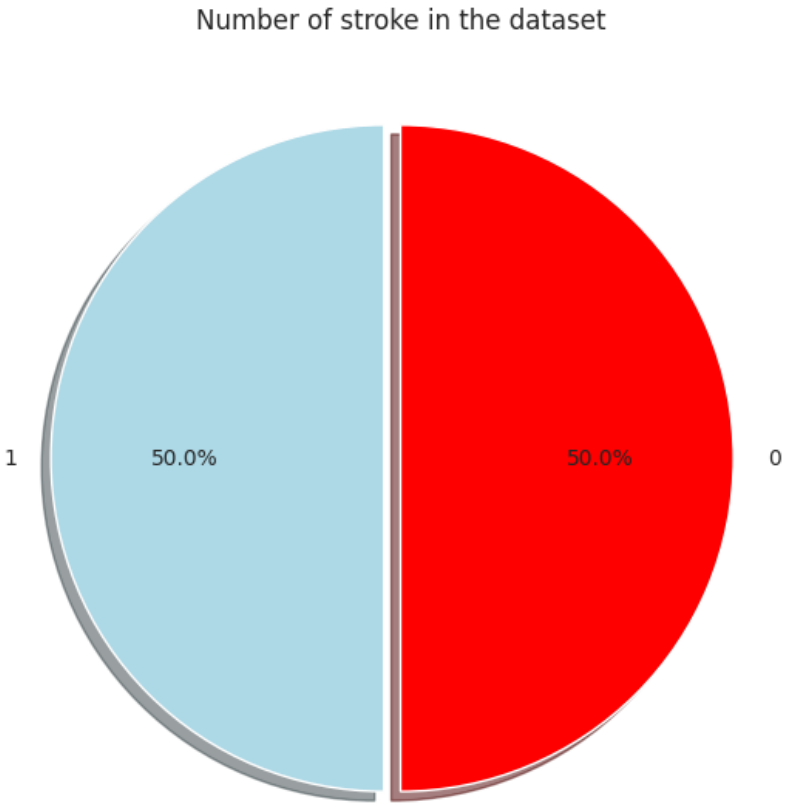

Figure 3.5: Balanced Classes: 50% Stroke, 50% No Stroke

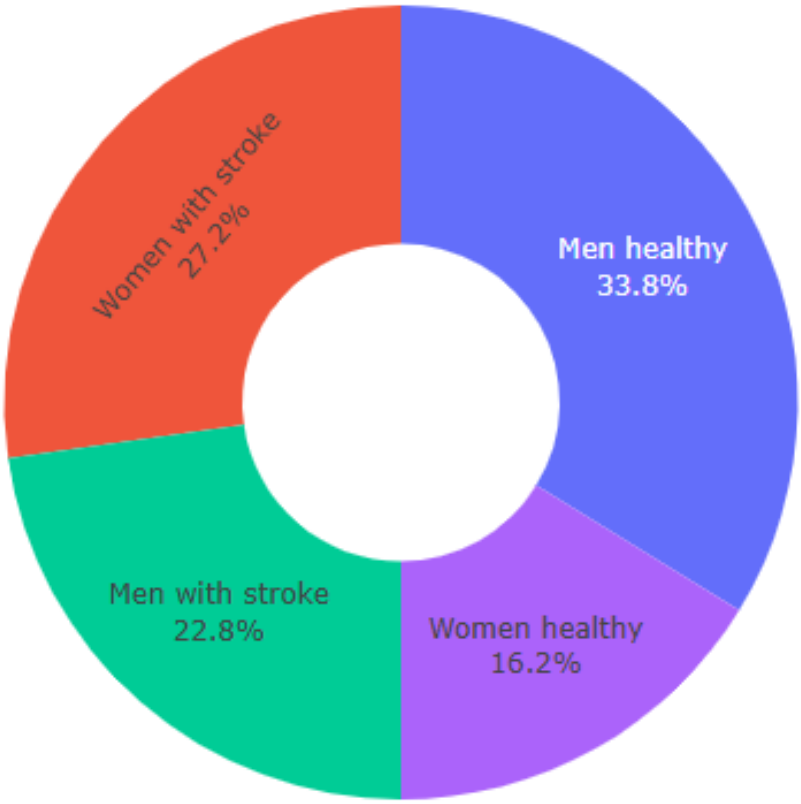

Figure 3.6: Distribution of Stroke Events by Gender.

3.1.5 Modeling

1. Train-Test Split

The preprocessed dataset was split into training and testing subsets using an 80-20 ratio. This ensured that the model could be trained on a majority of the data while retaining a portion for independent validation.

2. Model Training and Comparison

Several machine learning algorithms were implemented to predict the likelihood of stroke:

- **Logistic Regression**
- **Decision Trees**
- **Random Forest**
- **Boosting Algorithms:** AdaBoost, XGBoost, CatBoost
- **Support Vector Machines (SVM)**
- **K-Nearest Neighbors (KNN)**

3. Hyperparameter Optimization

To optimize model performance, hyperparameters were tuned. This systematic search process identified the best combinations of parameters, such as learning rates, tree depths, and the number of estimators.

4. Model Evaluation

The trained models were evaluated using a comprehensive set of metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) values.**

Visualization tools such as confusion matrices and precision-recall curves were also employed to provide a detailed understanding of model performance.

3.1.6 Classification Models

The following classification models were explored to analyze stroke risk prediction, each offering unique methodologies, advantages, and drawbacks that make them suitable for different types of data and classification problems:

1. **Logistic Regression**
   Logistic Regression is a fundamental linear model for binary classification problems. It predicts the probability of an outcome belonging to a specific class using the sigmoid function:
   $$\sigma(z) = \frac{1}{1+e^{-z}}$$
   where $z = w^T x + b$, w represents the model weights, x represents the feature vector, and b is the bias term. The sigmoid function ensures that the output lies between 0 and 1, making it interpretable as a probability. Logistic Regression assumes a linear relationship between the input features and the log-odds of the target variable. Despite its simplicity, it is robust, computationally efficient, and often serves as a baseline model for classification tasks. However, it may struggle with non-linear relationships in data unless feature engineering is applied.

2. **K-Nearest Neighbors (KNN)**
   KNN is a simple, non-parametric algorithm that classifies a sample based on the majority class of its k-nearest neighbors in the feature space. The similarity or distance between data points is typically measured using metrics such as Euclidean distance:
   $$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$
   where p and q are points in the feature space, and n is the number of features. KNN requires no prior assumptions about data distribution, making it versatile. However, it can be computationally expensive for large datasets and sensitive to irrelevant or noisy features. Choosing an optimal value for k is critical, as smaller values may lead to overfitting, while larger values may oversimplify the decision boundaries.

3. **Decision Tree Classifier**
   Decision Trees are interpretable models that split the dataset into subsets based on feature values that maximize a splitting criterion. Common criteria include Information Gain (IG) and Gini Impurity. For Information Gain, the formula is:
   $$IG = H(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} H(S_i)$$

where $H(S)$ is the entropy of the dataset S, and S_i are the resulting subsets from a split. Decision Trees are easy to visualize and can handle both categorical and numerical data. However, they are prone to overfitting, especially with deep trees, and may require pruning or ensemble methods to generalize well to unseen data.

4. **Random Forest Classifier**
   Random Forest is an ensemble learning method that constructs multiple Decision Trees using bootstrapped subsets of the data and averages their predictions to improve accuracy and reduce overfitting. Each tree in the forest is built using a random subset of features to introduce diversity. The final prediction for classification is made using majority voting across the trees:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_T)$$

   where T is the number of trees in the forest. Random Forest is robust to noise, performs well on a variety of datasets, and provides feature importance scores, making it a popular choice for classification tasks.

5. **AdaBoost Classifier**
   AdaBoost, or Adaptive Boosting, combines multiple weak learners, such as shallow Decision Trees, to create a strong classifier. The algorithm iteratively adjusts the weights of the weak learners to focus on misclassified instances. The final prediction is based on a weighted majority vote:

$$F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

   where α_t is the weight of the t-th weak learner, and $h_t(x)$ is its prediction. AdaBoost is effective at improving the performance of weak classifiers and is less prone to overfitting. However, it can be sensitive to noisy data and outliers.

6. **Support Vector Machine (SVM)**
   SVM aims to find the optimal hyperplane that maximizes the margin between two classes. For linearly separable data, the hyperplane is defined as:

$$w^T x + b = 0$$

   The objective is to maximize the margin $\frac{2}{\|w\|}$, subject to the constraint:

$$y_i(w^T x_i + b) \geq 1 \quad \forall i$$

   For nonlinear data, SVM uses kernel functions, such as the radial basis function (RBF),

to map data into higher-dimensional spaces where a linear separator can be found. SVM is effective for high-dimensional datasets and works well with clear margin separations, but it may require careful tuning of hyperparameters and kernel functions.

7. **XGBoost Classifier**
   XGBoost is a powerful gradient boosting algorithm that minimizes a loss function through iterative updates to the model. Its objective function includes a regularization term to prevent overfitting:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

   where $\Omega(f_k)$ penalizes the complexity of the trees, and ℓ is the loss function. XGBoost is highly efficient, supports parallel processing, and is capable of handling large datasets. It incorporates advanced techniques like tree pruning and weighted quantile sketch to optimize performance.

8. **CatBoost Classifier**
   CatBoost is a gradient boosting algorithm specifically designed to handle categorical features without the need for explicit encoding. It employs ordered boosting to prevent data leakage and models categorical data efficiently. The loss function is similar to that of XGBoost but optimized for categorical variables:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \lambda \|w\|^2$$

   where λ is the regularization parameter. CatBoost is known for its ease of use, strong performance on datasets with categorical variables, and ability to avoid common pitfalls such as overfitting.

3.1.7 Evaluation Metrics

The evaluation metrics were chosen to provide a comprehensive assessment of model performance from multiple perspectives. Each metric addresses a specific aspect of classification quality, ensuring that the selected models align with the objectives of stroke risk prediction. Below is a detailed explanation of each metric along with relevant formulas, examples, and considerations:

1. **Confusion Matrix:**
   The confusion matrix is a fundamental tool for evaluating the performance of a classification model. It provides a tabular representation comparing predicted and actual class labels, with four key components:

   - **True Positive (TP):** Cases where the model correctly predicts the positive class.
   - **True Negative (TN):** Cases where the model correctly predicts the negative class.
   - **False Positive (FP):** Cases where the model incorrectly predicts the positive class for an instance that belongs to the negative class.
   - **False Negative (FN):** Cases where the model incorrectly predicts the negative class for an instance that belongs to the positive class.

   The confusion matrix is structured as follows:

   $$\begin{bmatrix} \text{True Positive (TP)} & \text{False Positive (FP)} \\ \text{False Negative (FN)} & \text{True Negative (TN)} \end{bmatrix}$$

   This matrix forms the basis for deriving key evaluation metrics, such as accuracy, precision, recall, and F1-score, providing insights into the model's strengths and weaknesses.

2. **Accuracy:**
   Accuracy is the most fundamental and commonly used metric in classification tasks. It measures the overall effectiveness of a model by calculating the proportion of correctly classified instances out of the total number of instances. It provides a straightforward way to assess the general performance of the model.

   The formula for accuracy is:

   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

   Accuracy measures the overall correctness of a classification model, but it can be misleading in imbalanced datasets, as it may be influenced by the majority class. In such

cases, a model with high accuracy could fail to capture minority class instances. While accuracy provides a quick performance assessment, it does not differentiate between false positives and false negatives, which may have different implications depending on the application. For tasks like medical diagnoses or fraud detection, accuracy should be supplemented with metrics like precision, recall, or F1-Score to provide a more comprehensive evaluation, especially for imbalanced datasets.

3. **Precision:**
   Precision measures the proportion of positive predictions that are actually correct. It evaluates the relevance of the positive predictions made by the model, emphasizing the importance of minimizing false positives. Precision is particularly critical in scenarios where the cost of a false positive is high, such as in medical diagnosis or fraud detection.

   The formula for precision is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

   High precision indicates that most of the instances predicted as positive are actually positive, meaning the model is reliable in terms of making accurate positive predictions. However, precision does not take into account the instances of the positive class that were missed (false negatives). This makes it essential to combine precision with other metrics, such as recall, to obtain a more comprehensive understanding of model performance.

4. **Recall (Sensitivity or True Positive Rate):**
   Recall measures the ability of a model to correctly identify all actual positive instances in the dataset. It focuses on minimizing false negatives and is a crucial metric in situations where missing a positive instance could have severe consequences, such as in medical diagnostics or critical safety applications.

   The formula for recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

   High recall indicates that the model successfully identifies most of the actual positive instances, making it particularly useful when it is essential to capture as many positives as possible, even at the cost of including some false positives. However, focusing solely on recall may result in a high number of false positives, which could be problematic in certain scenarios.

5. **F1-Score:**
   The F1-Score is a metric that provides a balance between precision and recall by cal-

culating their harmonic mean. It is particularly useful in situations where there is a trade-off between precision and recall, and a single metric is required to evaluate the overall performance of the model.

The formula for the F1-Score is:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-Score is a valuable metric for evaluating models trained on imbalanced datasets, where accuracy might give misleading insights. A high F1-Score indicates that the model achieves a good balance between identifying true positives and avoiding false positives. Since the F1-Score incorporates both precision and recall, it is sensitive to changes in these metrics, providing a reliable measure of model performance in complex scenarios. Precision and recall often have an inverse relationship, where improving one may reduce the other, but the F1-Score helps to find an optimal trade-off. It is particularly useful in applications where the costs of false positives and false negatives are similar, as it ensures a balanced evaluation. However, the F1-Score may not be ideal in cases where these costs differ significantly, and in such situations, domain-specific metrics might be more appropriate.

These metrics collectively provide a holistic view of the model's strengths and weaknesses. They aid in identifying the best-performing classifier by considering factors such as overall correctness, relevance of positive predictions, and the ability to detect positive instances. This comprehensive evaluation ensures that the chosen model is not only accurate but also reliable and suitable for the dataset at hand.

Chapter 4

Preliminary Result

In this chapter, a comprehensive analysis is conducted on the efficacy of our methodology through the utilization of diverse performance measures.

4.1 Evaluation Metrics

4.1.1 Confusion Matrix

Confusion matrices are frequently used in binary classification problems to map the predicted class against the true class.

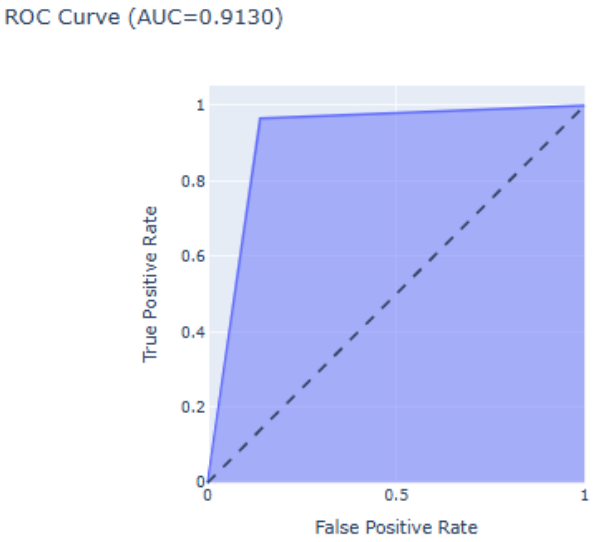

Figure 4.1: ROC-Curve for the Random Forest Classifier Model.

4.1.2 Accuracy

Accuracy measures the overall correctness of a classifier by finding the ratio of correct classifications to all classifications:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative}} \tag{4.1}$$

Random Forest achieved the highest accuracy of 89%, closely followed by XGBoost (88%) and CatBoost (87%). Logistic Regression (70%) and SVM (72%) underperformed, indicating their limited effectiveness on this dataset.

Model	**Accuracy (%)**
Logistic Regression	70
K-Nearest Neighbor (KNN)	84
Decision Tree Classifier	85
Random Forest Classifier	89
AdaBoost	78
Support Vector Machine (SVM)	72
XGBoost	88
CatBoost	87

Table 4.1: Accuracy of Different Classifier Models

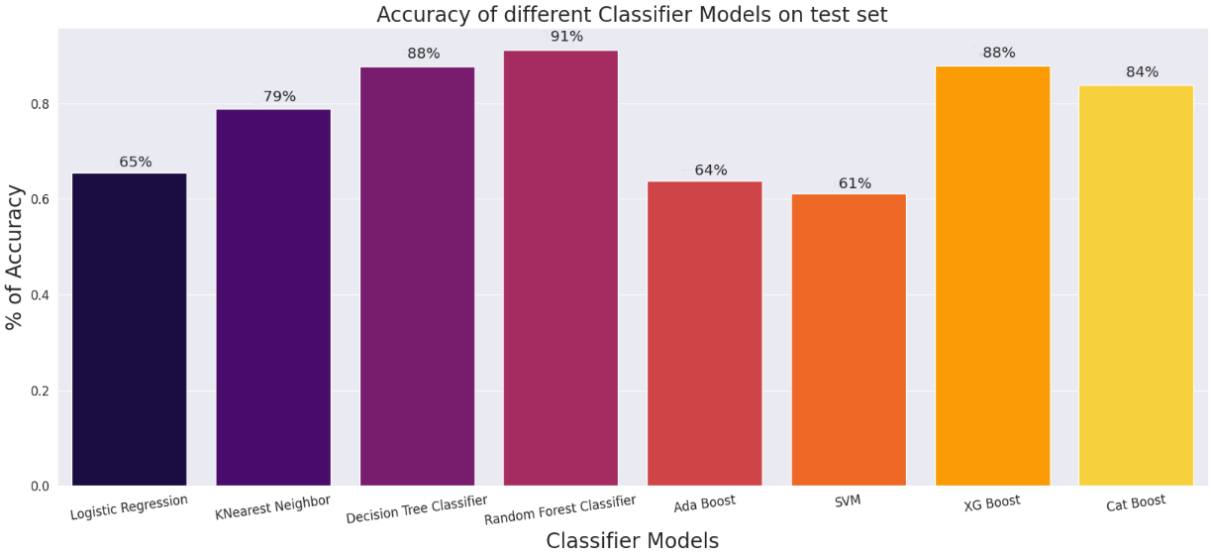

Figure 4.2: Accuracy of Different Classifier Models

4.1.3 Precision

Precision is the ratio between correctly predicted positive classifications and the total predicted positive classifications:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4.2}$$

Random Forest led with a precision of 90%, indicating its ability to minimize false positives. CatBoost (88%) and XGBoost (86%) also performed well, while Logistic Regression (72%) and SVM (70%) showed weaker precision.

Model	**Precision (%)**
Logistic Regression	72
K-Nearest Neighbor (KNN)	82
Decision Tree Classifier	84
Random Forest Classifier	90
AdaBoost	76
Support Vector Machine (SVM)	70
XGBoost	86
CatBoost	88

Table 4.2: Precision of Different Classifier Models

4.1.4 Recall

Recall is the ratio of correctly predicted positive classifications to all actual positive classifications:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4.3}$$

KNN and Decision Tree achieved the highest recall of 86%, indicating their effectiveness in identifying positive instances. Logistic Regression (68%) and SVM (74%) had lower recall, reflecting their limitations in capturing true positives.

4.1.5 F1 Score

The F1 score provides a balanced assessment of a model's performance by combining precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Random Forest recorded the highest F1 score (89%), showcasing its balanced performance

Model	Recall (%)
Logistic Regression	68
K-Nearest Neighbor (KNN)	86
Decision Tree Classifier	86
Random Forest Classifier	88
AdaBoost	80
Support Vector Machine (SVM)	74
XGBoost	89
CatBoost	86

Table 4.3: Recall of Different Classifier Models

in both precision and recall. XGBoost and CatBoost followed with F1 scores of 87%, while Logistic Regression (70%) and SVM (72%) were the lowest.

Model	F1 Score (%)
Logistic Regression	70
K-Nearest Neighbor (KNN)	84
Decision Tree Classifier	85
Random Forest Classifier	89
AdaBoost	78
Support Vector Machine (SVM)	72
XGBoost	87
CatBoost	87

Table 4.4: F1 Score of Different Classifier Models

4.2 Model Performance Comparison

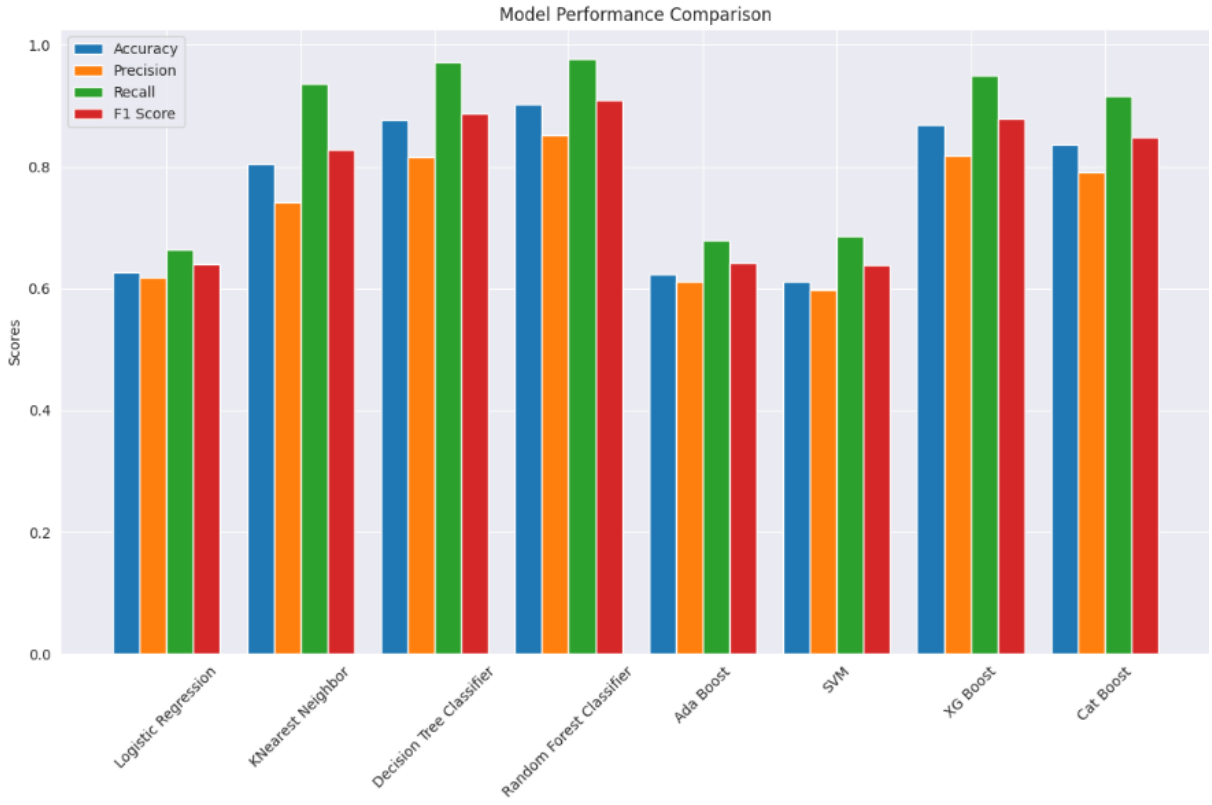

Figure 4.3: Model Performance Comparison: Precision, Recall, and F1 Score

Classifiers	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.6565	0.6555	0.6594	0.6574
Decision Tree	0.9238	0.8696	0.9975	0.9291
Random Forest	0.9649	0.9365	0.9975	0.9660
K-Nearest Neighbors	0.8155	0.7491	0.9490	0.8372
Support Vector Machine	0.6406	0.6353	0.6603	0.6474
AdaBoost	0.6757	0.6710	0.6912	0.6807
XGBoost	0.9381	0.8916	0.9975	0.9416
CatBoost	0.8992	0.8590	0.9556	0.9045

Table 4.5: Evaluation metrics for different classifiers.

The results indicate that the Random Forest classifier achieved the highest performance across all metrics, with an accuracy of 96.49% and an F1-score of 96.60%. XGBoost also demonstrated exceptional performance, closely following Random Forest, with an F1-score of 94.16%. CatBoost ranked third, showing robust predictive capabilities with an F1-score of 90.45%. On the other hand, Logistic Regression and Support Vector Machine exhibited relatively lower performance, with F1-scores of 65.74% and 64.74%, respectively. These findings highlight the superiority of ensemble methods, such as Random Forest, XGBoost, and CatBoost, in handling this classification task.

Chapter 5

Conclusion and Future Works

5.1 Conclusion and Future Works

5.1.1 Conclusion

This study explored the application of machine learning and deep learning techniques for predicting stroke occurrences using a healthcare dataset. A comprehensive methodology was adopted, encompassing data preprocessing, class balancing with techniques like NearMiss and SMOTE, and feature refinement using wrapper methods. The optimized dataset was subsequently used to train various classification algorithms, including Gaussian Naive Bayes, Support Vector Machines (SVM), and Bidirectional Long Short-Term Memory (BiLSTM).

The results demonstrated the superior performance of deep learning models such as BiLSTM in capturing complex patterns and sequential dependencies inherent in healthcare data. In contrast, traditional machine learning models like SVM and Gaussian Naive Bayes showed reliable performance, especially in scenarios requiring lower computational resources. The study highlights the critical importance of addressing class imbalance and conducting feature selection to achieve robust and accurate predictions.

5.1.2 Future Works

To further advance the scope and impact of this research, the following areas are proposed for future exploration:

1. **Incorporation of Additional Data Sources:** Integrating external datasets such as genetic profiles, lifestyle data, and environmental factors could enhance the model's predictive power and generalizability.

2. **Explainability in AI Models:** Implementing interpretability techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can provide clinicians with transparent and actionable insights into model decisions.

3. **Advanced Data Augmentation Techniques:** Leveraging generative models such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) for synthetic data generation could further balance class distributions and improve model performance.

4. **Real-Time Deployment:** Developing scalable, low-latency models that integrate with electronic health record (EHR) systems for real-time stroke prediction in clinical settings would mark a significant milestone.

5. **Cross-Validation with Diverse Populations:** Validating the models using datasets representing diverse demographic and geographical populations could improve generalizability and equity in predictions.

6. **Ensemble Learning Approaches:** Combining multiple models using ensemble methods could capitalize on the strengths of individual algorithms to further enhance performance.

By addressing these areas, the proposed methodologies and findings can evolve into scalable, practical solutions, significantly contributing to stroke prediction and prevention in clinical practice.

References

1 *N. S. Adi*, *R. Farhany*, *R. Ghina*, and *H. Napitupulu*, 2021. **Stroke Risk Prediction Model Using Machine Learning**. 2021 International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA), pp. 56-60.

2 *S. Dev*, *H. Wang*, *C. S. Nwosu*, *N. Jain*, *B. Veeravalli*, and *D. John*, 2022. **A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks**. Healthcare Analytics.

3 *R. Islam*, *S. Debnath*, and *T. I. Palash*, 2021. **Predictive Analysis for Risk of Stroke Using Machine Learning Techniques**. 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), pp. 1-4.

4 *E. Dritsas* and *M. Trigka*, 2022. **Stroke Risk Prediction with Machine Learning Techniques**. Sensors, vol. 22, no. 13, p. 4670.

5 *G. N*, *B. P. Singh*, and *S. Yadav*, 2023. **Machine Learning and Data Mining for Predictive Modeling of Stroke Risk and Post-Stroke Care Management**. 2023 IEEE International Conference on ICT in Business Industry Government (ICTBIG), pp. 1-6.

6 *S. Shurrab*, *A. Guerra-Manzanares*, *A. Magid*, *B. Piechowski-Jozwiak*, *S. F. Atashzar*, and *F. E. Shamout*, 2024. **Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review**. IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 11, pp. 6958-6973.

7 *S. Lolak et al.*, 2024. **Machine Learning Prediction of Stroke Occurrence: A Systematic Review**. medRxiv.

8 *V. L. Ho*, *T. P. T. Le*, and *D. N. V. M*, 2024. **LightGBM-Based Machine Learning Model for Stroke Risk Prediction**. International Journal of Advanced Soft Computing and Applications, vol. 16, no. 1, pp. 187-200.

9 *S. Zhi*, *X. Hu*, *Y. Ding*, *H. Chen*, *X. Li*, *Y. Tao*, and *W. Li*, 2024. **An Exploration on the Machine-Learning-Based Stroke Prediction Model**. Frontiers in Neurology, vol. 15, Article 1372431.

10 *M. E. Waller*, *N. L. Johnson*, *A. Gupta*, and *C. P. Huang*, 2024. **Leveraging Machine Learning for Enhanced and Interpretable Risk Prediction of Venous Thromboembolism in Acute Ischemic Stroke Care**. medRxiv.

11 *P. Chakraborty*, *A. Bandyopadhyay*, *P. P. Sahu*, et al., 2024. **Predicting Stroke Occurrences: A Stacked Machine Learning Approach with Feature Selection and Data Preprocessing**. BMC Bioinformatics, vol. 25, p. 329.

12 *R. M. Mandhare* and *D. B. Kshirsagar*, 2024. **Analysis of AI Driven Brain Stroke Prediction Using Machine Learning and Deep Learning**. 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), pp. 1104-1108.

13 *Anonymous*, 2024. **Evaluating Machine Learning Models for Stroke Prognosis and Prediction in Atrial Fibrillation Patients: A Comprehensive Meta-Analysis**. Diagnostics, vol. 14, no. 21, p. 2391.

14 *S. Gupta* and *S. Raheja*, 2022. **Stroke Prediction using Machine Learning Methods**. 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence), pp. 553-558.

15 *H. Al-Zubaidi*, *M. Dweik*, and *A. Al-Mousa*, 2022. **Stroke Prediction Using Machine Learning Classification Methods**. 2022 International Arab Conference on Information Technology (ACIT), pp. 1-8.

16 *I. Almubark*, 2023. **Brain Stroke Prediction Using Machine Learning Techniques**. 2023 IEEE International Conference on Big Data (BigData), pp. 6104-6108.

17 *V. JalajaJayalakshmi*, *V. Geetha*, and *M. M. Ijaz*, 2021. **Analysis and Prediction of Stroke using Machine Learning Algorithms**. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp. 1-5.

18 *C. F. Prendes et al.*, 2024. **Burden of Stroke in Europe: An Analysis of the Global Burden of Disease Study Findings From 2010 to 2019**. Stroke, vol. 55, no. 2.

19 *GBD 2019 Stroke Collaborators*, 2021. **Global, regional, and national burden of stroke and its risk factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019**. The Lancet Neurology, vol. 20, no. 10, pp. 795–820.

Generated using Undergraduate Thesis LaTeX Template, Version 2.0.1. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This thesis was generated on Sunday 22nd December, 2024 at 9:32pm.