Advanced Stroke Prediction Using Machine Learning and Deep Learning Techniques

by

Osama Elkott

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (MSc) in Computational Science

The Office of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

Abstract

Stroke remains a leading cause of disability and mortality, necessitating reliable predictive models for early diagnosis. This thesis enhances stroke prediction using a comprehensive Kaggle dataset of 43,400 clinical records, employing both machine learning and deep learning techniques. The methodology includes extensive data preprocessing, handling missing values, data imputation, encoding categorical variables, normalization, and addressing class imbalances, followed by the implementation and evaluation of algorithms such as Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, and Deep Neural Networks (DNNs). Metrics like accuracy, precision, recall, F1-score, AUC, and specificity were used for evaluation. Ensemble methods like Gradient Boosting (G-Boost) and Extreme Gradient Boosting (XGB) outperformed others, with G-Boost achieving an accuracy of 99% using a 70-20-10 split ratio. DNN also showed promise in handling complex data. The contributions highlight the potential of advanced techniques in stroke prediction, achieving superior performance and providing insights into health parameters influencing stroke risk. This study contributes to medical informatics by setting benchmarks for stroke prediction and demonstrating AI's applicability in healthcare. Implementing these models in clinical settings can enhance early detection, enable personalized interventions, and improve patient outcomes, reducing the burden of stroke.

Keywords: Stroke prediction, machine learning, deep learning, clinical data, Kaggle dataset, ADASYN method, Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Deep Neural Networks, Gradient Boosting, Extreme Gradient Boosting, medical informatics.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Kalpdrum Passi, for his invaluable guidance, continuous support, and encouragement throughout the course of my research. His expertise and insights have been instrumental in shaping this thesis. Dr. Passi's unwavering commitment to excellence and his profound knowledge have been a cornerstone of my academic journey, providing me with the inspiration and direction needed to overcome challenges and achieve my goals.

I would also like to thank my former supervisor, Dr. Amr Abdel-dayem, for his initial guidance and support during the early stages of this research. Dr. Amr's mentorship laid a strong foundation for this work, and his early contributions were pivotal in setting the trajectory for my research endeavors.

I extend my appreciation to the faculty and staff of the Laurentian University for their assistance and resources, which greatly facilitated my research. Their willingness to provide access to essential resources, coupled with their administrative support, has been critical to the successful completion of this thesis.

I am also grateful to my colleagues and friends, particularly, for their support that motivated me to push forward. Their feedback, discussions, and camaraderie have enriched my research experience and contributed to the development of this thesis.

I am indebted to my family for their unwavering support and understanding during this journey. My parents and my siblings have been a source of strength and comfort throughout this process. Their patience, belief in me, and emotional support have been instrumental in helping me stay focused and resilient.

Thank you all for your invaluable contributions. This thesis would not have been possible without your support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Stroke, a leading cause of death and disability globally, presents an immense challenge to public health, demanding urgent and effective interventions. Defined as a medical condition caused by the interruption of blood supply to the brain, stroke can result in long-lasting physical, cognitive, and emotional impairments, if not fatal outcomes. Recent statistics indicate that there are approximately 12 million new strokes each year globally, resulting in around 6.5 million deaths and another 5 million individuals left permanently disabled [1]. This places a significant burden on family and community care systems, as well as on healthcare resources [2][3].

The dichotomy of stroke, ischemic and hemorrhagic, with ischemic strokes accounting for about 87% of all cases, underscores the complexity of its prediction and prevention strategies [1] [4]. The general overview of stroke's global impact highlights not only its prevalence but also the variability of its incidence across different populations and regions. This variability is influenced by a range of risk factors, including hypertension, smoking, diabetes, obesity, physical inactivity, and unhealthy diet, many of which are modifiable with appropriate interventions [4]. The increasing prevalence of these risk factors, particularly in low and middle-income countries undergoing rapid urbanization and lifestyle changes, has led to a surge in stroke incidence, projecting an epidemiological transition that necessitates robust public health strategies [4].

The importance of stroke prediction lies in its potential to reduce morbidity and mortality. Early identification of individuals at high risk of stroke can facilitate the implementation of preventive measures, including lifestyle modifications and pharmacological interventions, to avert the onset

of stroke. Accurate prediction models can also aid in the efficient allocation of healthcare resources, directing preventive efforts towards those at highest risk and ensuring that interventions are both timely and cost-effective [1][4].

1.1 Types of Strokes

Stroke is primarily categorized into three types, each with unique causes, characteristics, and implications for treatment and prognosis. Understanding these distinctions is crucial for effective prevention, diagnosis, and management strategies [5].

1.1.1 Ischemic Stroke

Ischemic stroke, the most common type, accounts for about 87% of all stroke cases. It occurs when a blood clot obstructs a blood vessel supplying blood to the brain. The clot often forms in areas where arteries have been narrowed or blocked over time due to atherosclerosis, a condition characterized by the buildup of plaques inside the artery walls [1][3]. Ischemic strokes can be further subdivided into two main categories.

Thrombotic Stroke: This subtype happens when a blood clot (thrombus) forms in the arteries directly supplying blood to the brain. Thrombotic strokes are often associated with long-term artery damage due to high cholesterol, diabetes, and other conditions that affect blood vessels [3][6].

Embolic Stroke: In an embolic stroke, a clot or debris forms elsewhere in the body (commonly in the heart) and travels through the bloodstream to lodge in narrower brain arteries. This type is frequently seen in patients with heart conditions such as atrial fibrillation [3][6].

1.1.2 Hemorrhagic Stroke

Hemorrhagic stroke occurs when a blood vessel in the brain ruptures or leaks, leading to bleeding into or around the brain. This bleeding can increase pressure in the skull, causing damage to brain cells. Hemorrhagic strokes account for about 13% of stroke cases but are responsible for a higher mortality rate compared to ischemic strokes [1][6]. There are two primary forms:

Intracerebral Hemorrhage: This is the most common type of hemorrhagic stroke, occurring when an artery in the brain bursts, flooding the surrounding brain tissue with blood. High blood pressure is a leading cause, as it can weaken arteries over time [3][6].

Subarachnoid Hemorrhage: This less common form involves bleeding into the subarachnoid space between the brain and the membranes that cover it. It's often caused by the rupture of an aneurysm, a weak area in a blood vessel wall [6].

1.1.3 Transient Ischemic Attack (TIA)

Often referred to as a "mini-stroke" a TIA is a temporary blockage of blood flow to the brain, lasting less than 24 hours. While the symptoms of a TIA mirror those of a stroke, they typically resolve without permanent damage. However, TIAs are critical warning signs of a potential future stroke, and approximately a third of individuals who experience a TIA go on to have a stroke later on [6].

Understanding the types of strokes is fundamental for implementing targeted therapeutic interventions and preventive measures. For instance, treatment for ischemic stroke might involve medication to dissolve clots or prevent clot formation, while managing hemorrhagic stroke could

require surgery to repair damaged blood vessels. Moreover, recognizing and promptly addressing the signs of a TIA can significantly reduce the risk of a subsequent, more severe stroke [1][3][6].

Prevention strategies for all stroke types often focus on controlling risk factors, such as managing high blood pressure, diabetes, and cholesterol levels, quitting smoking, maintaining a healthy weight, engaging in regular physical activity, and following a balanced diet. Public health initiatives and individual education on recognizing stroke symptoms and seeking immediate medical attention can also significantly impact outcomes, underscoring the importance of awareness and proactive healthcare engagement in combating this global health issue [4][6].

1.2 Risk Factors

Stroke influenced by a complex interplay of various risk factors. These factors are broadly categorized into modifiable [7][8] and non-modifiable [9][8], highlighting the importance of lifestyle and medical interventions in stroke prevention. Understanding these risk factors is crucial for developing targeted strategies to reduce stroke incidence and improve public health outcomes.

1.2.1 Modifiable Risk Factors

Hypertension: Widely recognized as the primary risk factor for stroke, high blood pressure exerts excessive force against artery walls, leading to damage and increasing the risk of both ischemic and hemorrhagic strokes. Effective management of hypertension through lifestyle changes and medication can significantly reduce stroke risk.

Heart Diseases: Conditions such as atrial fibrillation (irregular heartbeat), coronary artery disease, and heart valve diseases markedly elevate the risk of stroke, particularly of the embolic ischemic type. Managing heart conditions and maintaining heart health are pivotal in stroke prevention.

Diabetes: This condition, characterized by high blood sugar levels, contributes to fatty deposits in blood vessels, facilitating clot formation and artery damage. Tight glucose control and diabetes management can mitigate the associated stroke risk.

Cholesterol Levels: High levels of LDL (bad) cholesterol and low levels of HDL (good) cholesterol can lead to the buildup of plaques in arteries (atherosclerosis), narrowing and potentially blocking these vessels, thereby increasing stroke risk.

Smoking: Smoking accelerates clot formation, thickens blood, and increases the buildup of plaques in arteries. Quitting smoking dramatically reduces the risk of stroke.

Physical Inactivity and Obesity: These factors contribute to the development of hypertension, diabetes, and high cholesterol levels, further elevating stroke risk. Regular physical activity and maintaining a healthy weight are essential preventive measures.

Poor Diet: Diets high in salt, saturated fats, and trans fats, and low in fruits, vegetables, and fiber can contribute to the development of stroke risk factors such as high blood pressure and atherosclerosis.

Excessive Alcohol Intake: Heavy drinking can lead to an increase in blood pressure and the risk of atrial fibrillation, both of which are significant stroke risk factors.

1.2.2 Non-Modifiable Risk Factors

Age: The risk of stroke increases with age, doubling for each decade of life after age 55. While age itself cannot be modified, understanding this risk can inform more rigorous monitoring and preventive measures in older adults.

Gender: Men have a higher risk of stroke than women at a younger age. However, women tend to have strokes at older ages and are more likely to die from stroke than men, partly due to longer life expectancy.

Family History and Genetics: Individuals with a family history of stroke or heart attack are at increased risk. Genetic factors can influence stroke risk through familial tendencies toward hypertension, diabetes, and other conditions.

Ethnicity: Certain ethnic groups, such as African Americans, have a higher risk of stroke than others. This increased risk is often due to higher prevalence rates of hypertension, diabetes, and obesity within these populations.

Previous Stroke or TIA: Individuals who have previously experienced a stroke or transient ischemic attack (TIA) are at a significantly higher risk of having another stroke.

Understanding and addressing these risk factors through public health initiatives, individual lifestyle changes, and medical interventions can significantly reduce the incidence of stroke. Healthcare providers play a crucial role in educating patients about their specific risks and the steps they can take to mitigate them, emphasizing the importance of a proactive approach to stroke prevention.

1.3 Current Stroke Management

Current stroke management encompasses a multifaceted approach aimed at rapid diagnosis, timely treatment, and comprehensive rehabilitation to minimize brain damage, preserve functionality, and prevent recurrence. Advances in medical science and technology over recent years have significantly improved stroke outcomes, emphasizing the importance of the "time is brain" principle in acute care. Stroke management can be broadly categorized into acute management, secondary prevention, and rehabilitation.

1.3.1 Acute Management

Rapid Diagnosis: The initial step in managing a stroke involves prompt recognition of symptoms, both by the public and healthcare providers. Tools like the FAST (Face drooping, Arm weakness, Speech difficulties, Time to call emergency services) test are crucial for early identification. Upon arrival at the hospital, immediate assessment using clinical evaluations and imaging tests, such as CT scans or MRIs, is essential to differentiate between ischemic and hemorrhagic stroke types, as the treatment varies significantly between them [10][11].

Treatment of Ischemic Stroke: For ischemic strokes, thrombolytic therapy using medications like alteplase (tPA) is the gold standard if administered within 4.5 hours of symptom onset. This treatment works by dissolving the blood clot obstructing the cerebral artery. Endovascular procedures, including mechanical thrombectomy, have extended the treatment window up to 24 hours for certain patients, showing remarkable success in removing large clots and restoring blood flow [10][12].

Management of Hemorrhagic Stroke: Treatment focuses on controlling bleeding and reducing pressure in the brain. This may involve surgical interventions to repair damaged blood vessels or remove accumulated blood. Medications to control blood pressure, prevent seizures, and reduce brain swelling are also commonly employed [10][12].

1.3.2 Secondary Prevention

After initial treatment, secondary prevention strategies are crucial to reduce the risk of subsequent strokes. This involves addressing modifiable risk factors and may include:

Anticoagulant Therapy: For patients with atrial fibrillation or other conditions that increase the risk of clot formation, anticoagulants are prescribed to prevent future ischemic strokes.

Blood Pressure Management: Hypertension being a leading risk factor for stroke, managing blood pressure through lifestyle changes and medication is paramount.

Lipid-lowering Therapy: Statins and other lipid-lowering medications are used to control high cholesterol levels, reducing the risk of plaque buildup in arteries.

Lifestyle Modifications: Dietary changes increased physical activity, smoking cessation, and moderation of alcohol intake are recommended to lower stroke risk [11][12].

1.4 Rehabilitation

Rehabilitation begins as soon as the patient is stable, often within days of the stroke. The goal is to help the patient regain as much independence as possible by improving physical, cognitive, and emotional functions. Rehabilitation is a personalized process, involving a multidisciplinary team

including physiotherapists, occupational therapists, speech and language therapists, and psychologists. Key components of stroke rehabilitation include:

Physical Therapy: Focuses on improving motor skills, balance, and coordination, helping patients relearn movements and activities of daily living.

Speech Therapy: Assists in recovering communication skills and addressing swallowing difficulties.

Occupational Therapy: Aims to enhance self-care skills and adapt living environments to the patient's needs.

Psychological Support: Addresses emotional and cognitive challenges, offering strategies to cope with changes in mental functions and mood.

Technological Advances and Research

Ongoing research and technological innovations continue to refine stroke management strategies. Developments in artificial intelligence and machine learning are improving diagnostic accuracy and predicting outcomes, while novel therapeutic agents and interventions are under investigation to extend treatment windows and enhance recovery [13][14][15].

In conclusion, current stroke management strategies emphasize a rapid, coordinated approach to acute care, targeted secondary prevention to reduce recurrence risk, and personalized rehabilitation to maximize recovery. Continued advancements in research and technology hold promise for even better outcomes in the future, offering hope to stroke patients and their families.

1.5 Time Sensitivity in Stroke Management

Time sensitivity is a critical factor in stroke management, encapsulated in the medical adage "time is brain." This phrase underscores the urgent need for prompt treatment following a stroke, as the longer the brain is deprived of blood flow, the greater the extent of neuronal damage and the poorer the prognosis for recovery. The importance of timely intervention in stroke management cannot be overstated, with every minute counting towards the preservation of brain function and the reduction of long-term disability [16][17].

The Golden Hour

The concept of the "golden hour" refers to the critical window of time following the onset of stroke symptoms during which medical interventions can significantly improve outcomes. For ischemic strokes, treatments such as thrombolysis (the administration of clot-busting drugs) are most effective when administered within 4.5 hours of symptom onset. Similarly, mechanical thrombectomy, a procedure to remove a clot obstructing blood flow to the brain, has been shown to be beneficial up to 24 hours after onset in certain cases, with the highest success rates observed when performed as soon as possible [10][17].

Pathophysiological Basis

The urgency of timely stroke treatment is rooted in the pathophysiology of stroke. When a part of the brain loses its blood supply, it immediately ceases to function properly. Neurons in the affected area begin to die within minutes due to the lack of oxygen and nutrients. However, surrounding the core of dead tissue is a region known as the ischemic penumbra, which consists of neurons that are at risk but can potentially be saved with rapid restoration of blood flow. Delay in treatment

leads to the expansion of the core area, increasing the extent of permanent damage and reducing the likelihood of functional recovery [10][17].

Public Awareness and Response

A significant barrier to timely stroke management is the delay in recognizing stroke symptoms and seeking medical attention. Public education campaigns have aimed to increase awareness of the signs of stroke, such as facial drooping, arm weakness, and speech difficulties, and to emphasize the importance of calling emergency services immediately. The faster a stroke is diagnosed and treated, the better the chances are for a favorable outcome, making public knowledge and response time critical components of stroke care [17].

Pre-hospital and In-hospital Strategies

Efforts to reduce time to treatment encompass both pre-hospital and in-hospital strategies. Pre-hospital strategies include the training of emergency medical services (EMS) personnel to recognize stroke symptoms, prioritize stroke patients for rapid transport, and notify receiving hospitals in advance to expedite care upon arrival. In-hospital strategies focus on streamlining the stroke care pathway, from the emergency department through imaging and treatment. This involves the establishment of stroke units, protocols for rapid assessment and imaging, and the availability of specialized treatment options around the clock [10][17].

Telemedicine and Mobile Stroke Units

Advancements in telemedicine and the deployment of mobile stroke units (MSUs) represent innovative approaches to reducing treatment delays. Telemedicine allows for the remote evaluation of patients by stroke specialists, facilitating quicker decision-making regarding

treatment. MSUs, ambulances equipped with CT scanners and telemedicine capabilities, enable the diagnosis and initiation of treatment for stroke patients before they reach the hospital, further narrowing the critical time window for intervention [16][17].

In conclusion, time sensitivity in stroke management is a fundamental principle that drives the entire spectrum of stroke care, from public education and EMS response to hospital protocols and innovative treatment modalities. The collective goal of these efforts is to minimize the time from stroke onset to treatment, maximizing the potential for recovery and minimizing the long-term impact of stroke on individuals and society. The adage "time is brain" serves as a constant reminder of the urgency and precision required in the fight against stroke, highlighting the race against time to save brain cells and improve outcomes for stroke patients worldwide.

1.6 Relevance to Field

The field of computational science and medicine plays a pivotal role in enhancing stroke prediction methodologies. With advancements in data collection and processing capabilities, researchers now have access to vast amounts of health data, including electronic health records, genetic information, and lifestyle data, which can be analyzed to identify patterns and risk factors associated with stroke. Machine learning and deep learning algorithms, subsets of artificial intelligence (AI), offer promising tools for analyzing these data sets, enabling the development of predictive models that are both accurate and personalized. These advanced techniques can handle complex, non-linear relationships between risk factors and stroke outcomes, providing valuable insights into stroke prediction.

Machine learning models, including decision trees, random forest, and support vector machines, have been applied to stroke prediction with varying degrees of success. These models are adept at identifying patterns within data, making them useful for handling structured datasets. Deep learning models, such as deep neural networks (DNNs), further enhance predictive capabilities by automatically extracting high-level features from raw data, particularly useful in image analysis and complex datasets.

The relevance of stroke prediction research within computational science and medicine extends beyond the development of prediction models. It encompasses the integration of these models into clinical practice, where they can support decision-making processes and improve patient outcomes. Implementing machine learning-based prediction tools in healthcare settings requires careful consideration of their interpretability, reliability, and ethical implications, ensuring that they complement, rather than replace, clinical judgment.

Furthermore, the field of computational science contributes to the understanding of stroke mechanisms and the identification of novel biomarkers for risk stratification. By analyzing genetic data and molecular pathways, researchers can uncover biological markers that indicate an increased risk of stroke, providing targets for preventive interventions and personalized medicine approaches.

In conclusion, the prediction and prevention of stroke represent critical areas of research within computational science and medicine, addressing a global health challenge of significant proportions. The integration of machine learning and data analytics into stroke research offers the potential to revolutionize stroke prediction, enabling the development of models that are both sophisticated and clinically applicable. As the field advances, the focus will increasingly shift

towards the ethical and practical aspects of implementing these technologies in healthcare, ensuring that they contribute to the reduction of stroke incidence and the improvement of patient outcomes worldwide. This research endeavor not only aligns with the goals of public health and personalized medicine but also highlights the interdisciplinary nature of combating stroke, involving contributions from computational science, clinical medicine, public health, and beyond.

1.7 Machine learning

Machine learning is a specialized field within artificial intelligence (AI) that focuses on creating algorithms and statistical models. These models allow computer systems to enhance their performance on specific tasks by learning from experience. Unlike traditional programming, where tasks are explicitly defined, machine learning systems learn and improve from data, recognizing patterns and making decisions without direct programming [18]. Machine learning finds applications in various domains, such as data analysis, image and speech recognition, natural language processing, recommendation systems, and autonomous vehicles. It has become an integral part of technological advancements and continues to play a crucial role in the development of intelligent systems [18].

Types of Machine Learning

1. **Supervised Learning:** In supervised learning, algorithms are trained on labeled datasets, meaning each input comes with a corresponding output. The model learns to map inputs to outputs by analyzing the training data and then uses this understanding to predict outcomes for new, unseen data. This type of learning is widely used in applications such as spam detection, image classification, and medical diagnosis.

2. **Unsupervised Learning:** Unlike supervised learning, unsupervised learning involves training algorithms on data without labeled responses. The system attempts to identify patterns, structures, or relationships within the dataset. Common applications of unsupervised learning include clustering, where the algorithm groups similar data points together, and dimensionality reduction, which simplifies data without losing significant information. This approach is valuable in exploratory data analysis and customer segmentation.

3. **Reinforcement Learning:** Reinforcement learning is characterized by an algorithm learning to make decisions by performing certain actions and receiving feedback from its environment in the form of rewards or penalties. The objective is to develop a strategy that maximizes cumulative rewards over time. This type of learning is extensively used in robotics, game playing, and autonomous driving, where the system must learn complex behaviors through interaction.

1.8 Deep Learning

Deep learning, a subset of machine learning, involves neural networks with many layers that can learn increasingly abstract representations of the input data. This allows deep learning models to handle large-scale and complex datasets effectively, making significant strides in areas like image and speech recognition, natural language processing, and more. Deep learning has become a pivotal part of AI advancements, contributing to the development of sophisticated systems that mimic human-like understanding and decision-making processes [67].

Types of Deep Learning

1. **Deep Neural Networks (DNNs)**: DNNs consist of multiple layers of neurons that can capture complex patterns in data. Each layer extracts higher-level features from the output

of the previous layer. DNNs are particularly effective in tasks such as speech recognition, image classification, and medical diagnosis, where they can learn intricate representations of the input data.

2. **Convolutional Neural Networks (CNNs)**: CNNs are specialized for processing grid-like data structures, such as images. They use convolutional layers to automatically detect spatial hierarchies of features, making them highly effective for image and video analysis. In medical imaging, CNNs can identify patterns indicative of diseases, such as tumors or lesions, which are critical for diagnosis and treatment planning.

3. **Recurrent Neural Networks (RNNs)**: RNNs are designed to handle sequential data, making them ideal for tasks involving time-series data or natural language processing. They have the capability to remember previous inputs through their internal state, enabling them to model temporal dependencies. RNNs are used in applications such as language translation, speech recognition, and financial forecasting.

1.9 Stroke Prediction and Prevention

Stroke prediction and prevention constitute pivotal aspects of contemporary healthcare strategies aimed at mitigating the incidence and impact of one of the most formidable challenges in public health. With strokes being a leading cause of disability and mortality worldwide, the ability to accurately predict and effectively prevent strokes can dramatically alter the landscape of healthcare outcomes, significantly reducing the burden on patients, families, and healthcare systems.

1.10 The Role of Prediction in Stroke Prevention

The cornerstone of stroke prevention lies in the identification of individuals at high risk through predictive modeling. Recent advancements in computational science, data analytics, and artificial intelligence have ushered in a new era of precision medicine, where stroke prediction models harness vast arrays of data—from genetic markers to lifestyle factors—to forecast stroke risk with unprecedented accuracy.

These models integrate traditional risk factors, such as hypertension, diabetes, and smoking, with novel biomarkers and social determinants of health, offering a holistic view of an individual's stroke risk. Machine learning and Deep learning algorithms, capable of analyzing complex, non-linear relationships within large datasets, play a crucial role in refining these predictive models, enabling them to adapt and improve over time with the incorporation of new data.

Implementing Preventive Measures

Equipped with knowledge from predictive models, healthcare providers can implement targeted preventive strategies tailored to an individual's risk profile. This personalized approach to prevention may include pharmacological interventions, such as antihypertensive medications, anticoagulants for individuals with atrial fibrillation, and statins for managing cholesterol levels. Additionally, lifestyle modifications play a critical role in stroke prevention, with evidence-based recommendations emphasizing the importance of a healthy diet, regular physical activity, smoking cessation, and moderation of alcohol consumption.

Public health initiatives focusing on education and awareness are fundamental in promoting these lifestyle changes. Campaigns designed to inform the public about the signs of stroke and the

significance of modifiable risk factors empower individuals to take proactive steps towards stroke prevention.

1.11 Challenges and Ethical Considerations

Despite the progress in stroke prediction and prevention, challenges remain. Disparities in healthcare access and quality, socioeconomic factors, and the prevalence of risk factors in certain populations contribute to unequal stroke risks and outcomes. Addressing these disparities requires a concerted effort from policymakers, healthcare providers, and communities to implement inclusive, equitable healthcare strategies and social interventions.

Moreover, the dynamic nature of stroke risk—wherein an individual's risk profile may change over time—necessitates continuous monitoring and adjustment of preventive measures. The integration of predictive models into clinical practice and the development of user-friendly digital health tools for risk assessment and monitoring represent opportunities to enhance the effectiveness of stroke prevention strategies.

Despite its potential, the integration of machine learning or deep learning into clinical practice presents challenges. Data privacy and security are paramount concerns, given the sensitive nature of health information. Furthermore, the decisions are made through complex, often opaque processes—raises ethical questions regarding interpretability and accountability. Ensuring that these models are explainable and their decisions understandable to clinicians is crucial for their ethical use in healthcare.

Moreover, the reliance on large, annotated datasets for training machine learning and deep learning models introduces potential biases, particularly if the data is not representative of the broader

population. Efforts to mitigate these biases and ensure the generalizability of machine learning and deep learning applications are essential for their fair and effective implementation.

1.12 Future Directions

The future of stroke prediction and prevention is promising, with ongoing research focused on identifying new risk factors, refining predictive algorithms, and exploring innovative preventive interventions. The potential of emerging technologies, such as wearable devices for real-time health monitoring and telemedicine for expanding access to preventive care, is particularly exciting. These advancements promise to make stroke prevention more accessible, personalized, and effective, ultimately leading to a reduction in stroke incidence and an improvement in public health outcomes.

In conclusion, stroke prediction and prevention are integral to reducing the global burden of stroke. Through the use of advanced predictive models, targeted preventive strategies, and public health initiatives, there is significant potential to decrease the incidence of stroke and improve the quality of life for individuals at risk. As technology and research continue to evolve, the prospects for further advancements in stroke prevention are both hopeful and boundless, signaling a brighter future in the fight against this devastating condition.

Rehabilitation and Recovery

In the realm of stroke rehabilitation, machine learning and deep learning models contribute to the development of personalized rehabilitation programs. By analyzing data from wearable sensors, mobile health applications, and patient-reported outcomes, these models can track progress, adjust exercises, and even predict recovery trajectories. Additionally, machine learning -powered virtual

reality and robotic devices are emerging as innovative tools for enhancing motor recovery, providing engaging, adaptive, and effective rehabilitation experiences.

Looking Ahead

The future of machine learning in stroke care is promising, with ongoing research and development focused on refining predictive models, enhancing diagnostic tools, and innovating treatment and rehabilitation strategies. Collaborations between computer scientists, clinicians, and patients are key to realizing the full potential of machine learning in transforming stroke care. As technology advances, so too will the capabilities of machine learning and deep learning models, promising a future where stroke prediction, diagnosis, and treatment are more accurate, personalized, and effective than ever before.

In summary, the role of machine learning and deep learning in stroke care marks a significant advancement in our ability to predict, diagnose, and treat this complex condition. By harnessing the power of machine learning, healthcare professionals can offer more targeted and effective care, improving outcomes for stroke patients and moving closer to a future where the burden of stroke is significantly reduced.

1.13 Objectives

In this thesis, the focus is on addressing a significant gap in the methodologies for predicting stroke, leveraging a comprehensive dataset from Kaggle containing 43,400 records [69]. This Study aims to:

1- Develop and apply multiple machine learning and deep learning techniques to analyze complex health data for stroke prediction.
2- Achieve superior performance in stroke prediction compared to existing models by focusing on various metrices such as accuracy, precision, recall, AUC, F1-score, and specificity.
3- Provide more effective tool for early detection and intervention of strokes, thereby reducing stroke incidence and improving patient care through tailored risk managements strategies.
4- Contribute to the field of medical informatics by advancing the use of AI and machine learning and deep learning in healthcare, setting new benchmark for future research in stroke prediction.

Addressing these objectives is significant for multiple reasons. Enhancing stroke prediction accuracy directly benefits healthcare by enabling early identification and intervention, which can significantly reduce the risk of severe outcomes and improve patient recovery rates. For patients, more accurate predictions mean personalized risk assessments, leading to targeted prevention strategies and minimizing unnecessary treatments. Furthermore, this research enriches the field of machine learning and deep learning in medicine by showcasing the applicability of ML and DL models in analyzing complex health datasets, contributing valuable insights and methodologies that can be adapted for other medical conditions.

1.14 Contribution

This study makes the following contributions:

1- **Dataset:** This study distinguishes itself by leveraging a comprehensive and recent dataset sourced from Kaggle that includes 12 features. it's one of the most extensive and up-to-date datasets in current research.

2- **Enhanced Machine Learning and Deep learning Capability:** This research introduces a pioneering approach by applying machine learning and deep learning techniques to clinical datasets for predictive purposes, specifically targeting Stroke.

3- **Accuracy and Results:** In this study, the machine learning models demonstrated an outstanding accuracy rate of 99%. This achievement signifies a remarkable improvement compared to prior research endeavors. The substantial increase in accuracy is a testament to the efficacy of the methodologies employed and the significance of utilizing a comprehensive dataset.

The primary focus of this heightened accuracy was directed towards predicting patients who might be at risk of suffering from Stroke.

The exceptional accuracy achieved in this study not only marks a substantial improvement in predictive capabilities but also holds profound implications for advancing the field and enhancing the accuracy of Stroke prediction.

Previous studies have primarily focused on traditional statistical methods, clinical assessments, and basic demographic analyses for stroke prediction. This study pioneers the use of machine

learning for stroke prediction, which can handle large datasets and identify non-linear relationships, thereby enhancing predictive accuracy.

1.15 Thesis outline

This thesis is organized into six chapters, each serving a specific purpose in the overall narrative of the research. Here is a brief overview of what each chapter entails:

Chapter 2: Literature Review - This chapter provides a comprehensive review of the existing literature in the fields of Stroke prediction, and machine learning. It explores previous research on the use of clinical datasets with machine learning in Stroke Prediction, highlighting the current state of knowledge, and identifying gaps that this study aims to address.

Chapter 3: Data and Pre-processing - This chapter extensively explores the initial phases of data collection, and pre-processing.

The chapter delves into the pre-processing procedures implemented to guarantee the data's quality and pertinence for subsequent analysis, encompassing techniques for handling missing data, and imbalanced classes.

Chapter 4: Methods– This chapter provides a detailed exploration of the methodology applied to the dataset prepared in chapter 3. The section outlines specific techniques employed in this process. Additionally, it elucidates the procedures involved in selecting and training machine learning and deep learning models, along with the metrics used for evaluating model performance. This chapter offers an in-depth examination of the analytical methods employed throughout the study.

Chapter 5: Results and Discussion - This chapter presents the results obtained from the machine learning and deep learning models. It provides a detailed analysis of these results, discussing them in the context of the study's objectives and the existing literature.

Chapter 6: Conclusion and Future Work - The final chapter concludes the thesis by summarizing the key findings of the study. It discusses the implications of these findings and their contribution to the field. It also suggests potential avenues for future research, building on the work done in this study.

By following this structure, the thesis aims to provide a comprehensive exploration of the potential of machine learning and deep learning in Stroke prediction.

Chapter 2

Literature Review

Advances in machine learning and deep learning have shown promise in enhancing the accuracy and reliability of stroke prediction models. This chapter reviews existing literature on stroke prediction, focusing on various machine learning and deep learning techniques, and their applications.

2.1 Background on Stroke Prediction

Stroke remains a paramount medical emergency, necessitating prompt and precise diagnosis to mitigate potentially devastating consequences. Over the past decade, a significant body of literature has emerged, reflecting collaborative efforts to deepen our understanding of stroke, refine diagnostic strategies, and improve patient management. This literature review endeavors to comprehensively survey and synthesize the extensive research conducted on stroke prediction, encompassing diagnostic models, predictive tools, machine learning applications, and evolving methodologies.

Stroke, often referred to as a "medical catastrophe" due to its sudden onset and varied clinical presentation, demands innovative approaches for identification, prediction, and management. This review meticulously examines seminal studies, navigating the landscape of stroke research to unveil the evolution of diagnostic criteria, the integration of advanced predictive models, and the adoption of cutting-edge machine learning techniques.

A multidisciplinary approach involving clinical, radiological, and computational methodologies emerges as essential in unraveling the complexities of stroke. In subsequent sections, this exploration delves into diverse studies shaping the contemporary landscape of stroke research, addressing critical aspects like risk prediction, diagnostic accuracy, and the integration of machine learning. Through this synthesis, the aim is to consolidate existing knowledge, identify gaps, and pave the way for future research endeavors aimed at refining the approach to diagnosing and managing stroke. Table 2.1 succinctly summarizes key studies in stroke research, showcasing diverse methodologies and advancements in diagnostic and predictive approaches.

Table 0.1: Related Work Summary

Ref	Dataset	Preprocessing	Methods	Performance
Rahman et al. [19]	Dataset: Kaggle; # of features: 12; Sample size: 5110	Imputation: Most frequent value Encoding: Label encoding Balancing: Random oversampling Scaling: Min-Max scaling Dimensionality Reduction: PCA	3 ANN layers 4 ANN layers LR DT RF KNN SVM GNB BNB XGBoost Adaboost LGBM	ACC: 0.840; AUC: 0.91; F1: 0.860 ACC: 0.923; AUC: 0.97; F1: 0.936 ACC: 0.71; AUC: 0.79; F1: 0.71 ACC: 0.98; AUC: 0.98; F1: 0.97 ACC: 0.99; AUC: 1.00; F1: 0.99 ACC: 0.96; AUC: 0.98; F1: 0.96 ACC: 0.82; F1: 0.81 ACC: 0.70; AUC: 0.78; F1: 0.70 ACC: 0.67; AUC: 0.71; F1: 0.63 ACC: 0.97; AUC: 0.98; F1: 0.97 ACC: 0.78; AUC: 0.76; F1: 0.78 ACC: 0.95; AUC: 0.96; F1: 0.95
Md. Ashrafuzzaan et al. [20]	Dataset: Kaggle; # of features: 12; Sample size: 5110	Imputation: Mean value Data Cleaning: Removed unnecessary data Encoding: Label encoding Shuffling: Dataset shuffled Normalization: Numerical & categorical normalization	LR DT RF SVM Naive Bayes CNN	ACC: 0.95; F1: 0.98 ACC: 0.926; F1: 0.96 ACC: 0.947; F1: 0.98 ACC: 0.95; F1: 0.98 ACC: 0.875; F1: 0.92 ACC: 0.955; F1: 0.98
Liu et al. [21]	Dataset: Kaggle; # of features: 12; Sample size: 43400	Outlier Removal: Age < 25 & BMI > 60 Normalization: Z-score normalization Imputation: Prediction model (RFR)	DNN with Automated Hyperparameter Optimization (AutoHPO), DNN Bagging RF XGB Ada Average scores	ACC: 0.71; Rec: 0.674 ACC: 0.651; Rec: 0.677 ACC: 0.738; Rec: 0.876 ACC: 0.728; Rec: 0.893 ACC: 0.741; Rec: 0.887 ACC: 0.726; Rec: 0.895 ACC: 0.733; Rec: 0.888
Kokkotis et al. [22]	Dataset: Kaggle; # of features: 12; Sample size: 43400	Outlier Removal: Age & BMI Normalization: Standard Scaler Balancing: Undersampling	LR RF XGB KNN SVM MLP	ACC: 0.735; Rec: 0.781 ACC: 0.711; Rec: 0.792 ACC: 0.725; Rec: 0.783 ACC: 0.691; Rec: 0.788 ACC: 0.712; Rec: 0.804 ACC: 0.708; Rec: 0.814
Jing et al. [23]	Dataset: Kaggle; # of features: 12; Sample size: 5110	Outlier Removal Imputation: Simple, regression imputation Encoding: Label, one-hot encoding	DNN 50 epochs 100 epochs 200 epochs	 ACC: 0.89; AUC: 0.57; F1: 0.22 ACC: 0.88; AUC: 0.55; F1: 0.17 ACC: 0.86; AUC: 0.53; F1: 0.15

		Balancing: SMOTE, undersampling Dimensionality Reduction: PCA-KMeans	DNN + focal loss 50 epochs 100 epochs 200 epochs	 ACC: 0.78; AUC: 0.62; F1: 0.25 ACC: 0.76; AUC: 0.68; F1: 0.29 ACC: 0.75; AUC: 0.72; F1: 0.31
			DNN + SMOTE + focal loss 50 epochs 100 epochs 200 epochs	 ACC: 0.80; AUC: 0.58; F1: 0.24 ACC: 0.78; AUC: 0.59; F1: 0.26 ACC: 0.75; AUC: 0.64; F1: 0.28
			PCA-kmeans + DNN + focal loss 50 epochs 100 epochs 200 epochs	 ACC: 0.86; AUC: 0.61; F1: 0.33 ACC: 0.71; AUC: 0.76; F1: 0.47 ACC: 0.92; AUC: 0.90; F1: 0.77
Hung et al. [24]	Dataset: NHIRD; # of features: 7,932; Sample size: 900,000	Imputation: Removed missing data Scaling: Feature values scaled to 0-1 Balancing: Undersampling	DNN GBDT LR SVM	ACC: 0.873; Rec: 0.845 ACC: 0.868; Rec: 0.856 ACC: 0.866; Rec: 0.820 ACC: 0.839; Rec: 0.813
Bacchi et al. [25]	Dataset: CT brain images Collected from 2 hospitals' database over 7 years	Imputation: Median imputation Scaling: Feature scaling	CNN ANN CNN CTB	ACC: 0.71; AUC: 0.70; Rec: 0.93; F1: 0.74 ACC: 0.68; AUC: 0.68; Rec: 0.43; F1: 0.55 ACC: 0.68; AUC: 0.63; Rec: 0.71; F1: 0.67
Cheon et al. [26]	Dataset: Korean National Hospital database; # of features: 11; Sample size: 15000	Imputation: Converted binary/categorical to continuous Dimensionality Reduction: PCA (standard, min/max, quantile) Class Weighting	DNN GNB KNNC SVC ADB RFC	ACC: 0.84; AUC: 0.834 ACC: 0.706; AUC: 0.78 ACC: 0.796; AUC: 0.721 ACC: 0.702; AUC: 0.715 ACC: 0.792; AUC: 0.792 ACC: 0.784; AUC: 0.775
Heo et al. [27]	Dataset: Demo-graphic Dataset; # of features: 38 Sample size: 2604	Data Cleaning: Deleted rows with missing values	DNN ASTRAL RF LR	AUC: 0.888 AUC: 0.839 AUC: 0.857 AUC: 0.849
Chen et al. [28]	Dataset: EHR database of three hospitals; # of features: 23	Data Integration: Multiple sources	SVM + SMOTE DT + SMOTE RF + SMOTE DNN + SMOTE DNN + GAN DNN + AIT	ACC: 0.731; AUC: 0.809; Rec: 0.545; F1: 0.658 ACC: 0.707; AUC: 0.767; Rec: 0.504; F1: 0.642 ACC: 0.715; AUC: 0.755; Rec: 0.495; F1: 0.645 ACC: 0.737; AUC: 0.818; Rec: 0.530; F1: 0.667 ACC: 0.725; AUC: 0.798; Rec: 0.612; F1: 0.691 ACC: 0.745; AUC: 0.819; Rec: 0.630; F1: 0.716

2.2 Literature Review of Stroke

The Framingham Stroke Risk Profile, developed by Wolf et al. [37], is a seminal work in stroke prediction, providing a comprehensive risk assessment tool based on data from the Framingham Heart Study. This model incorporates various risk factors such as age, blood pressure, diabetes, smoking status, cardiovascular disease, and atrial fibrillation to estimate the probability of stroke. The Framingham Stroke Risk Profile has been instrumental in highlighting the multifactorial nature of stroke risk and is widely used in clinical practice to guide prevention strategies.

The Cox proportional hazards model, as described by Kalbfleisch and Prentice [38], is a robust statistical method for analyzing time-to-event data, commonly used in medical research to assess the impact of various risk factors on stroke occurrence. This model allows for the inclusion of multiple covariates and provides estimates of the hazard ratios for different risk factors, making it a powerful tool for understanding the temporal dynamics of stroke risk. Its application in stroke research has helped elucidate the contributions of individual risk factors over time.

The SPARCL criteria, developed by Amarenco et al. [39], emerged from the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial. This study demonstrated the efficacy of high-dose atorvastatin in reducing the risk of recurrent stroke in patients with a history of stroke or transient ischemic attack. The SPARCL criteria have been pivotal in promoting the use of aggressive lipid-lowering strategies to prevent recurrent stroke, influencing treatment guidelines and clinical practice.

The CHADS2 score, introduced by Gage et al. [40], is a clinical tool used to predict the risk of stroke in patients with atrial fibrillation. This scoring system combines factors such as congestive heart failure, hypertension, age, diabetes, and prior stroke to estimate stroke risk. The CHADS2 score is widely adopted in clinical practice to stratify patients and guide anticoagulant therapy decisions, thereby preventing stroke in high-risk individuals.

Adams et al. [41] developed the TOAST classification system to categorize subtypes of acute ischemic stroke based on etiology. This system includes categories such as large artery atherosclerosis, cardioembolism, small vessel occlusion, stroke of other determined etiology, and stroke of undetermined etiology. The TOAST classification is widely used in both clinical and

research settings to improve the diagnosis and management of ischemic stroke, facilitating a standardized approach to stroke classification and treatment.

2.3 Literature Review of Stroke Prediction

In studies related to stroke prediction and its subsets, the key measure of success is primarily assessed through the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), along with sensitivity and specificity. Analyzing these parameters allows for a comprehensive statistical evaluation of the results in the below-mentioned research studies and facilitates meaningful comparisons between them.

The choice of methodology for predicting stroke is often influenced by the type of dataset sources available. Some studies utilize imaging techniques such as Computed Tomography (CT) Scans and Magnetic Resonance Imaging (MRI), while others combine clinical variables and CT images. As a result, researchers often adopt a hybrid approach, incorporating both clinical and imaging data. Additionally, some studies focus specifically on clinical measurements and demographic data, employing machine learning techniques. Various researchers utilize a diverse array of machine-learning strategies to analyze these datasets, including logistic regression, decision trees, random forests, support vector machines, and deep learning models.

Nielsen et al. [33] developed a deep learning model to predict tissue outcome and assess treatment effect in acute ischemic stroke using only imaging data from MRI scans. The study utilized diffusion-weighted imaging (DWI) and perfusion-weighted imaging (PWI) to create a predictive model capable of estimating the final infarct volume and predicting patient outcomes. The dataset comprised 320 patients with acute ischemic stroke, and the deep learning model achieved a high

accuracy in predicting tissue outcome and assessing the effectiveness of various treatment strategies. This study demonstrates the potential of using MRI data exclusively to enhance stroke diagnosis and treatment planning, providing a non-invasive method to support clinical decision-making.

Robben et al. [34] developed a deep learning model to predict final infarct volume from native CT perfusion (CTP) images and treatment parameters using a convolutional neural network (CNN). The study utilized a dataset comprising CT perfusion scans from patients with acute ischemic stroke, collected from multiple hospitals participating in the MR CLEAN Registry. The model achieved an AUC of 0.85, indicating high accuracy in predicting the final infarct volume. This study highlights the effectiveness of using deep learning models with CTP images to enhance the prediction of stroke outcomes, aiding in treatment planning and decision-making processes. By focusing solely on imaging data, the research underscores the potential of advanced neural networks to provide critical insights for stroke management.

Liu et al. [35] developed a deep convolutional neural network (CNN) to automatically segment acute ischemic stroke lesions using multi-modality MRI images. The study utilized a dataset consisting solely of MRI scans, including diffusion-weighted imaging (DWI), fluid-attenuated inversion recovery (FLAIR), and apparent diffusion coefficient (ADC) maps. The CNN model demonstrated high accuracy and robustness in lesion segmentation, achieving a Dice similarity coefficient of 0.81. This study highlights the potential of deep learning models in enhancing the precision and efficiency of stroke diagnosis by leveraging advanced imaging techniques. The exclusive use of MRI data underscores the capability of CNNs to analyze complex neuroimaging data for improved clinical outcomes.

Stier et al. [36] developed a deep learning model to predict tissue fate in acute ischemic stroke using CT and MRI imaging modalities. The study utilized imaging data from patients to create a predictive model that could identify tissue at risk of infarction based on imaging features. The model employed convolutional neural networks (CNNs) to analyze the imaging data, demonstrating a significant improvement in prediction accuracy over traditional methods. The study highlighted the potential of deep learning in enhancing the precision of stroke diagnosis and treatment planning by providing detailed insights into tissue fate and infarction risk.

The application of machine learning techniques has revolutionized stroke prediction. Logistic regression models have been traditionally used due to their simplicity and interpretability. However, more complex models such as random forests and support vector machines have shown improved predictive performance. Recently, deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated significant potential in analyzing imaging data and sequential clinical records, respectively.

Rahman et al. [19] conducted a study to predict brain stroke using various machine learning and deep neural network techniques on a dataset from Kaggle, consisting of 5110 samples. They utilized several classifiers including Extreme Gradient Boosting (XGBoost), AdaBoost, LightGBM, Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbors, SVM, Naive Bayes, and both 3-layer and 4-layer artificial neural networks (ANNs). The Random Forest classifier achieved the highest accuracy of 99%. The study demonstrated that machine learning techniques generally outperformed deep neural networks in stroke prediction, highlighting Random Forest's superior performance in predictive accuracy and efficiency.

Md. Ashrafuzzaman et al. [20] proposed a CNN model for stroke prediction, utilizing a healthcare dataset from Kaggle with 5110 samples, including features like age, gender, hypertension, heart disease, and BMI. They applied thorough data preprocessing steps and feature selection techniques such as univariate selection, feature importance, and correlation matrix analysis. Their 1D-CNN model achieved a high validation accuracy of 95.5%, significantly outperforming traditional machine learning models like Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine. The study underscores the critical role of deep learning in medical predictions, emphasizing its superior accuracy and ability to automatically identify significant features. This research demonstrates the potential for early and accurate stroke diagnosis using advanced CNN models, providing a promising tool for healthcare applications.

Liu et al. [21] developed a hybrid machine learning approach for predicting cerebral stroke using a large and imbalanced dataset with 43,400 records, including 783 stroke cases. Their method involved two main steps: imputing missing values with random forest regression and applying automated hyperparameter optimization (AutoHPO) with a deep neural network (DNN) for classification. The approach achieved notable results, including a false negative rate of 19.1%, a false positive rate of 33.1%, an accuracy of 71.6%, and a sensitivity of 67.4%. This method significantly reduces the false negative rate, enhancing stroke prediction reliability and offering a valuable tool for clinical decision-making.

Kokkotis et al. [22] developed an explainable machine learning pipeline for stroke prediction using imbalanced data, aiming to improve prediction reliability and interpretability. They compared six classifiers, finding the Multi-Layer Perceptron (MLP) achieved the lowest false-negative rate (18.60%). They used Shapley Additive Explanations (SHAP) to interpret model decisions, highlighting the impact of various risk factors on stroke prediction. This approach offers advanced,

effective risk stratification strategies for stroke patients, enabling timely diagnosis and appropriate treatments. The study demonstrates the potential of AI in enhancing stroke prediction and patient outcomes.

Jing et al. [23] conducted a performance analysis to predict stroke using an imbalanced medical dataset. They investigated potential stroke risk factors and applied four methods: ensemble weight voting classifier, SMOTE, PCA with K-Means Clustering, and DNN with Focal Loss. SMOTE and PCA-KMeans combined with DNN-Focal Loss significantly outperformed other methods, demonstrating superior performance on imbalanced datasets. The study highlights the effectiveness of these techniques in improving stroke prediction accuracy, especially for datasets with severe class imbalances.

Hung et al. [24] conducted a comprehensive study comparing the performance of deep neural networks (DNN) with other machine learning algorithms, such as gradient boosting decision trees (GBDT), logistic regression (LR), and support vector machines (SVM), for stroke prediction using a large-scale electronic medical claims (EMC) database of 800,000 patients. The study found that both DNN and GBDT achieved similarly high prediction accuracies, outperforming LR and SVM. Notably, DNN demonstrated optimal results using less patient data compared to GBDT, highlighting its efficiency and robustness in handling complex EMC data for accurate stroke prediction. The study underscores the potential of DNNs in clinical decision support systems for early stroke detection, emphasizing the importance of selecting appropriate machine learning algorithms to improve patient outcomes.

In their study, Cheon et al. [26] investigate the use of deep learning models to predict stroke patient mortality. The authors utilized data from the Korean National Hospital Discharge In-depth Injury

Survey (KNHDS) collected between 2013 and 2016, involving 15,099 stroke patients. They employed a deep neural network (DNN) approach combined with scaled principal component analysis (PCA) to extract relevant features from the data. This model incorporated variables such as gender, age, type of insurance, mode of admission, brain surgery status, hospital region, length of hospital stay, and the Charlson Comorbidity Index (CCI) score. Their DNN/PCA model achieved an AUC of 83.48%, outperforming several other machine learning methods including random forest and AdaBoost classifiers. This study highlights the potential of deep learning in enhancing the predictive accuracy of stroke outcomes using accessible medical service use and health behavior data. The findings emphasize the importance of early detection and personalized treatment plans to improve stroke prognosis and reduce healthcare costs.

Heo et al. [27] conducted a study to develop and compare machine learning models for predicting long-term outcomes in ischemic stroke patients. Utilizing a retrospective cohort of 2,604 patients from the Yonsei University Health System, they evaluated three models: deep neural networks (DNN), random forests, and logistic regression. The DNN model achieved the highest AUC at 0.888, significantly outperforming the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, which had an AUC of 0.839. This study demonstrates the superior performance of DNNs in predicting favorable outcomes, defined as a modified Rankin Scale (mRS) score of 0-2 at 3 months post-stroke, and underscores the potential of machine learning techniques to enhance prognostic accuracy in clinical settings. The findings highlight the efficacy of incorporating comprehensive clinical variables and advanced algorithms in improving stroke prognosis and patient management.

Chen et al. [28] proposed a novel Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) framework to address the challenges of small and imbalanced stroke datasets. The

study utilized data from multiple sources, including stroke data and data on chronic diseases like hypertension and diabetes, leveraging transfer learning to enhance prediction accuracy. They employed generative adversarial networks (GANs) for generating synthetic instances and Bayesian optimization for fine-tuning the model parameters. Their framework was validated on both synthetic and real-world data, demonstrating superior performance compared to existing state-of-the-art stroke risk prediction models. The results showed that HDTL-SRP achieved higher accuracy, recall, and F1 score, effectively balancing the stroke dataset and preserving patient privacy. This approach highlights the potential of integrating transfer learning and advanced optimization techniques in medical predictive modeling, providing a scalable solution for stroke risk prediction across multiple healthcare institutions.

Hybrid approaches that combine imaging data with clinical and demographic information have become increasingly popular. These methods leverage the strengths of both data types, leading to more robust and comprehensive predictive models. For example, integrating MRI findings with patient-specific risk factors such as age, gender, and medical history can provide a more nuanced understanding of stroke risk.

Through this synthesis of methodologies and findings, the aim is to consolidate existing knowledge, identify gaps, and pave the way for future research endeavors aimed at refining the approach to diagnosing and managing stroke. Table 2.1 succinctly summarizes key studies in stroke prediction research, showcasing diverse methodologies and advancements in diagnostic and predictive approaches.

Bacchi et al. [25] conducted a study utilizing deep learning to predict functional outcomes of ischemic stroke thrombolysis. They aimed to improve the accuracy of predictions regarding patient

recovery following thrombolysis treatment using imaging data and clinical features. The study compared the performance of deep learning models with traditional machine learning approaches. Results showed that deep learning models, particularly convolutional neural networks (CNNs), significantly outperformed conventional methods in predicting post-treatment outcomes. This highlights the potential of advanced AI techniques in enhancing clinical decision-making and personalized treatment strategies for stroke patients.

Park et al. [29] developed a 3D convolutional neural network (CNN)-based algorithm to predict mechanisms of acute ischemic stroke using brain MRI data, specifically diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) maps. The study involved 2,251 patients with acute ischemic stroke from Chungbuk National University Hospital. Their lesion segmentation model achieved a Dice score of 0.843, indicating high accuracy in identifying stroke lesions. For stroke subtype classification, the model reached an average accuracy of 81.9%, with specific accuracies of 81.6% for large artery atherosclerosis (LAA), 86.8% for cardioembolism (CE), and 72.9% for small vessel occlusion (SVO). This study underscores the potential of 3D-CNN models in providing accurate lesion segmentation and subtype classification, facilitating more precise diagnosis and treatment planning for stroke patients. By focusing on MRI data, their method offers a reliable tool for early stroke subtype classification, which is crucial for effective treatment and secondary prevention.

Karthik et al. [30] conducted a comprehensive review of recent advancements in deep learning techniques for brain stroke detection and lesion segmentation, highlighting the state-of-the-art methods and future prospects. The review analyzed 113 research papers from various academic databases, focusing on the use of Convolutional Neural Networks (CNNs) and Fully Convolutional Networks (FCNs) in stroke detection using CT and MRI imaging modalities. The study

emphasized the effectiveness of deep learning models in achieving accurate and automated stroke detection, discussing key challenges such as the need for high-resolution images and the diversity in brain tissues. The review also categorized different deep architectures based on their imaging modalities, providing insights into how these models have advanced stroke lesion detection and segmentation. This extensive survey concludes by identifying technical and non-technical challenges and suggesting future research directions to further enhance the application of deep learning in medical image analysis for stroke detection.

Xie et al. [31] conducted a study to predict stroke from electrocardiograms (ECGs) using a deep neural network approach. They developed a DenseNet-based classifier to analyze 12-lead ECG data, proposing it as a novel, data-driven method for stroke prediction without requiring expert domain knowledge. The study utilized ECG data from 100 subjects, equally divided between normal individuals and stroke patients, labeled based on CT images and clinical symptoms. Their model achieved a training accuracy of 99.99% and a prediction accuracy of 85.82%. This research highlights the potential of using ECG as a complementary diagnostic tool for stroke, demonstrating that deep learning can effectively map the nonlinear relationship between ECG patterns and stroke occurrence. The study's findings suggest that ECGs, typically used for cardiac monitoring, can also play a significant role in early stroke detection and diagnosis.

Hilbert et al. [32] explored data-efficient deep learning methods for predicting outcomes after endovascular treatment in acute ischemic stroke patients using CT angiography (CTA) images. The study included 1,301 patients from the MR CLEAN Registry and employed a Residual Neural Network (ResNet) with Structured Receptive Field Neural Networks (RFNN) and auto-encoder (AE) initialization for network weights. Their models, specifically the RFNN-ResNet-AE fine-tuned, achieved an average AUC of 0.71 for predicting good functional outcomes (mRS $\leq$ 2) and

0.65 for good reperfusion (mTICI ≥ 2b), outperforming traditional radiological image biomarkers. The study highlighted the effectiveness of using deep learning for automated image analysis, which does not require manual image annotation and is faster to compute. Model visualization techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), were used to interpret the decision-making process, revealing that arteries were significant features for predicting functional outcomes. This research underscores the potential of deep learning in improving stroke outcome predictions and treatment selection.

Chapter 3

Data and Preprocessing

This chapter details the datasets used for this research. The datasets form the foundation of the study, providing the necessary information for analyzing stroke prediction. This section will elaborate on the sources of data, the types of data included, and a thorough examination of each dataset's characteristics. Additionally, the chapter will explore the context and relevance of the data to stroke prediction research, ensuring a comprehensive understanding of the dataset's scope and limitations.

3.1 Dataset

Kaggle Stroke Prediction Dataset

Source: The Kaggle Stroke Prediction Dataset is a widely used dataset available on Kaggle, a platform for predictive modeling and analytics competitions. The data is curated from various healthcare records, providing a robust foundation for stroke prediction analysis [69].

Sample Size: The dataset comprises 43,400 samples, making it sufficiently large to allow for meaningful statistical analysis and machine learning and deep learning applications.

Features: The dataset includes 12 features: id, age, gender, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status, and stroke occurrence. Each feature plays a critical role in understanding the risk factors and predictors of stroke.

Data Type: The data is structured and primarily text-based, with numerical and categorical variables.

Context and Relevance: This dataset is crucial for developing predictive models for stroke because it includes a diverse set of variables that cover demographic, medical history, and lifestyle factors. These factors are often correlated with stroke risk, making the dataset valuable for building comprehensive predictive models.

3.2 Dataset structure

The dataset is characterized by 12 distinct features, each contributing valuable insights into the factors that may influence stroke risk. These features are as follows:

1. **id**: A unique identifier for each patient, ensuring the anonymity and confidentiality of the data.

2. **gender**: Categorized as "Male," "Female," or "Other," this feature recognizes the importance of gender differences in stroke risk.

3. **age**: The age of the patient, a critical factor in stroke prediction, given the varying risk across different age groups.

4. **hypertension**: Indicates with a binary value (0 for absence, 1 for presence) whether the patient has hypertension, a known risk factor for stroke.

5. **heart_disease**: Similar to hypertension, this binary feature (0 for absence, 1 for presence) identifies patients with heart disease.

6. **ever_married**: Captures marital status with "No" or "Yes" options, considering social and lifestyle factors in health outcomes.

7. **work_type**: Classifies the patient's employment into categories such as "children," "Govt_job," "Never_worked," "Private," or "Self-employed," reflecting the potential impact of occupational factors on health.

8. **residence_type**: Distinguishes between "Rural" and "Urban" living environments, acknowledging the environmental influences on health.

9. **avg_glucose_level**: Represents the average glucose level in the patient's blood, a key indicator of potential health issues related to stroke.

10. **bmi**: The body mass index, providing a measure of body fat based on height and weight.

11. **smoking_status**: Describes the patient's smoking habits with options like "formerly smoked," "never smoked," "smokes," or "unknown," recognizing the significant impact of smoking on stroke risk.

12. **stroke**: The target variable, with a binary value indicating whether the patient has experienced a stroke (1) or not (0).

3.3 Dataset representation

Dataset visualization section illuminates the underlying patterns and distributions within the stroke dataset, offering a graphical representation of the complex interplay between various demographic and health-related factors. These visualizations are crucial as they provide intuitive insights that might not be immediately evident through raw data analysis. By presenting the data in visual form,

we can observe trends, outliers, and correlations that inform the machine learning and deep learning models used for stroke prediction.

In this section, we explore a series of charts that illustrate the relationships between stroke incidence and key variables such as age, gender, average glucose level, and smoking status. Visual analytics serve not only to validate the integrity of the dataset but also to enhance our understanding of the stroke risk landscape. Through these visual narratives, the dataset's characteristics are dissected, providing a foundational understanding that supports the subsequent modeling and analysis efforts detailed in this thesis.

3.3.1 Gender distribution

Figure 3.1 visualizes the distribution of individuals by gender and their stroke status within the study's dataset. The x-axis categorizes participants into two groups: those who have not experienced a stroke (0) and those who have (1). The y-axis indicates the count of individuals within each category. Color coding distinguishes the genders: Male, Female, and Other. with the 'Other' category showing the least count, indicating either a smaller sample size or a lower stroke occurrence in this group. Notably, Female are more likely to have a stroke. This representation is critical in illustrating the gender dynamics within the stroke data, which may provide valuable insights into the gender-related risk factors associated with stroke incidence.

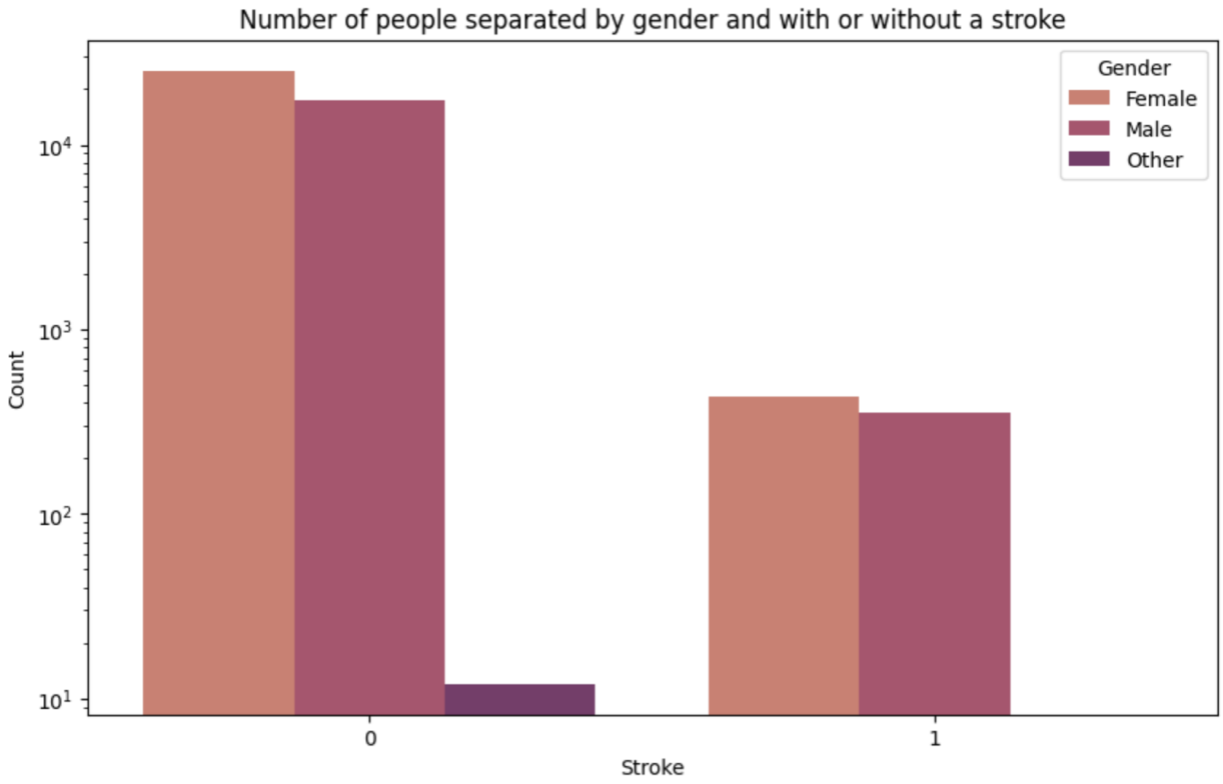

Figure 0.1: Gender distribution.

3.3.2 Age distribution

Figure 3.2 depicts the age distribution of individuals categorized by their stroke status. The x-axis differentiates between those who have not had a stroke (0) and those who have (1). The y-axis represents the count of individuals within each age group. The bars are color-coded to distinguish between age groups, which range from 0-25 years to 60-100 years. The chart suggests a higher concentration of younger individuals who have not experienced a stroke, while the incidence of stroke appears to increase with age, as shown by the larger counts in the higher age brackets among those who have had a stroke. The most prominent age group for stroke occurrence falls within the 60-100 year range, underscoring the increased stroke risk associated with advancing age.

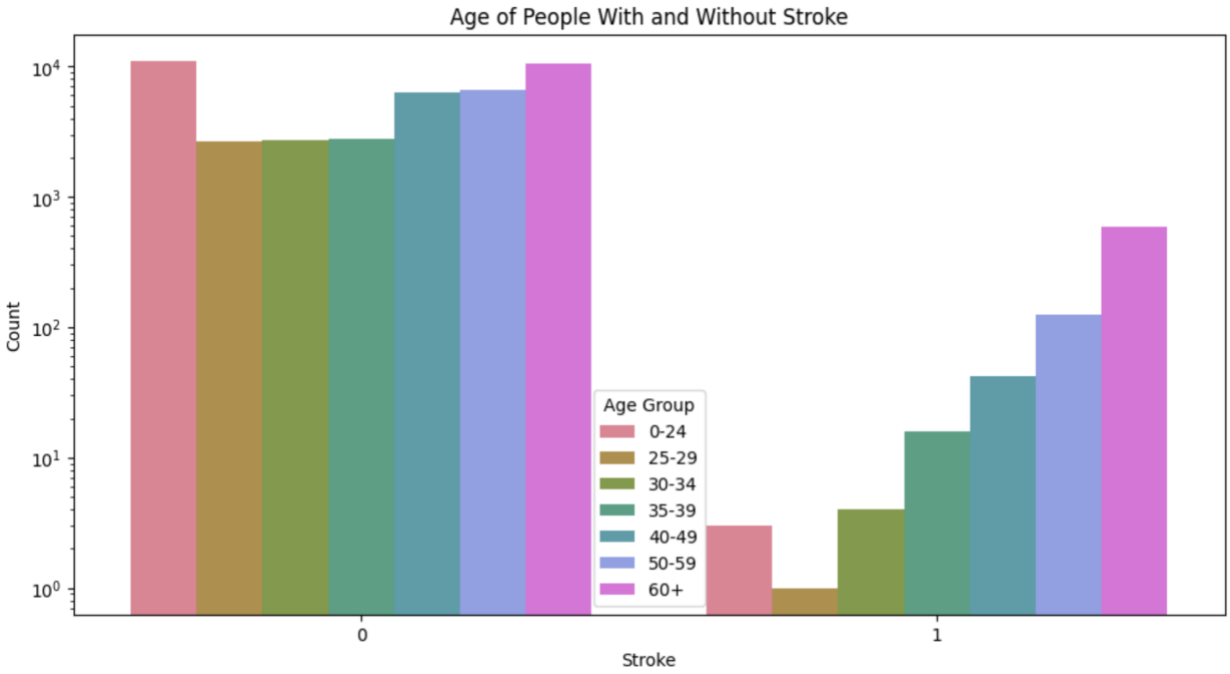

Figure 0.2: Age distribution.

3.3.3 Hypertension distribution

Figure 3.3 shows the proportion of stroke patients with and without hypertension. The chart displays the percentage of stroke patients categorized by their hypertension status. The larger segment represents patients without hypertension, while the smaller segment shows the proportion of patients with hypertension. This chart is an effective visual tool for quickly grasping the prevalence of hypertension among stroke patients in the dataset.

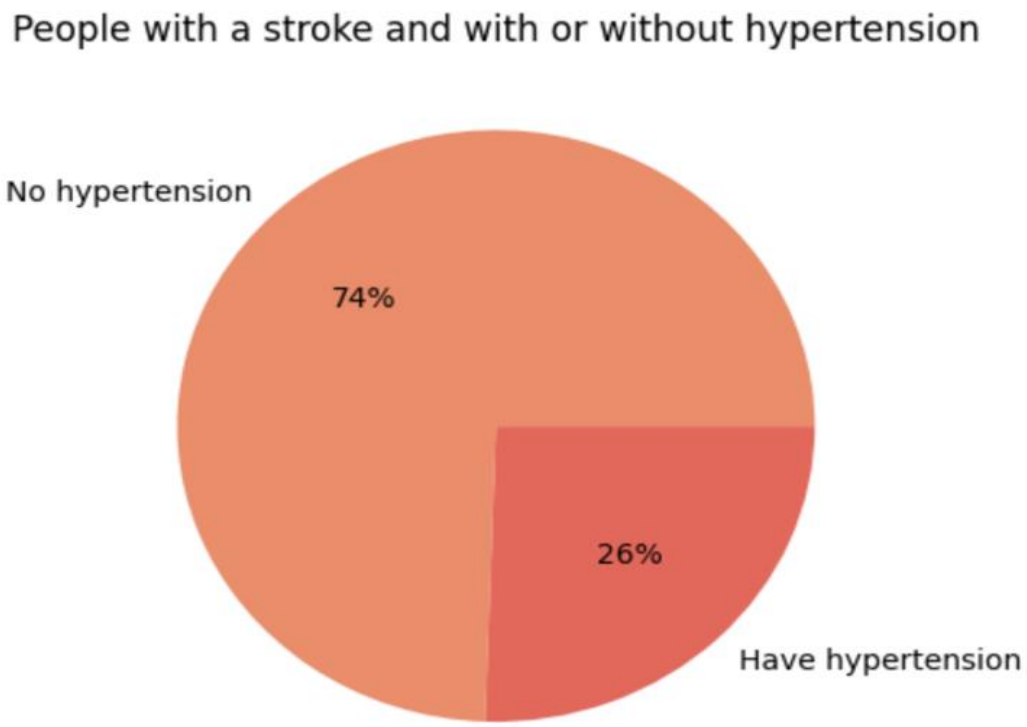

Figure 0.3: Stroke Patients with and Without Hypertension.

Figure 3.4 illustrates the average age of individuals with and without hypertension, separated by gender. The x-axis represents the hypertension status, with '0' indicating no hypertension and '1' indicating the presence of hypertension. The y-axis shows the average age of the individuals in each category. The bars are color-coded to differentiate between males and females. This visualization aids in understanding any age-related differences in hypertension prevalence among the genders within the study's population.

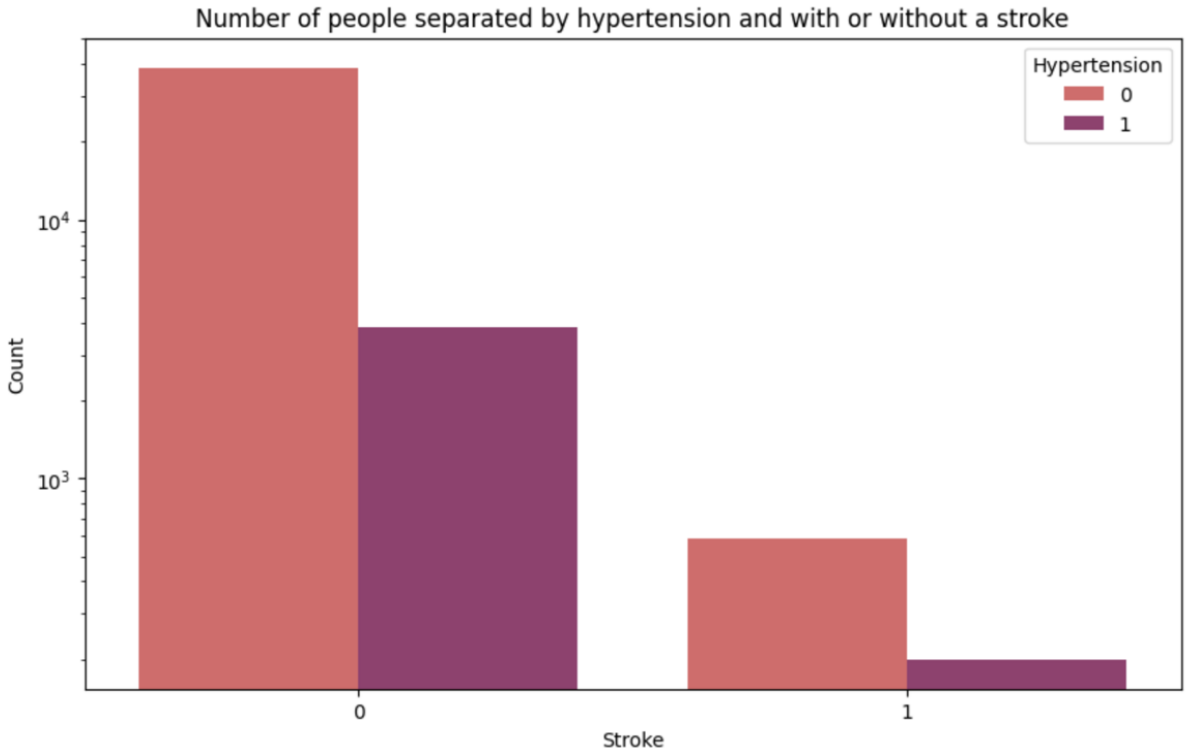

Figure 0.4: Hypertension distribution.

Figure 3.5 presents the number of people separated by their hypertension and stroke status and their age. The x-axis indicates hypertension status, with '0' for individuals without hypertension and '1' for those with hypertension. The y-axis shows the count of individuals in each category. The bars, distinguished by color, represent different ages. This graph highlights the relationship between stroke occurrence and hypertension within the study's cohort and demonstrates that age is not a significant factor for hypertension prevalence.

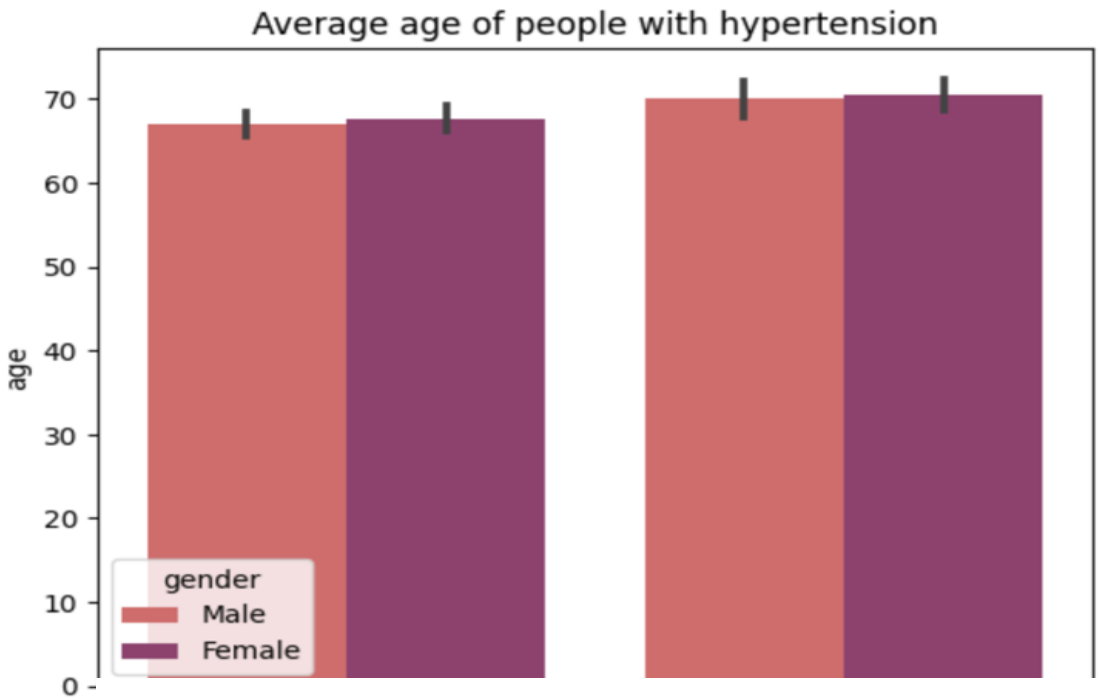

Figure 0.5: Average Age of People Categorized by Hypertension Status and age.

3.3.4 Heart disease distribution

Figure 3.6 depicts the relationship between heart disease and stroke occurrences within a population. The x-axis categorizes individuals based on stroke incidence: '0' for no stroke and '1' for stroke occurrence. The y-axis indicates the count of individuals. The bars are color-coded to represent heart disease status, with one color for individuals without heart disease ('0') and another for those with heart disease ('1'). The chart reveals a stark contrast in the number of individuals with and without heart disease in both stroke and non-stroke groups, providing insights into the correlation between these two health conditions. Showing that the heart disease does not has effective impact on stroke occurrence.

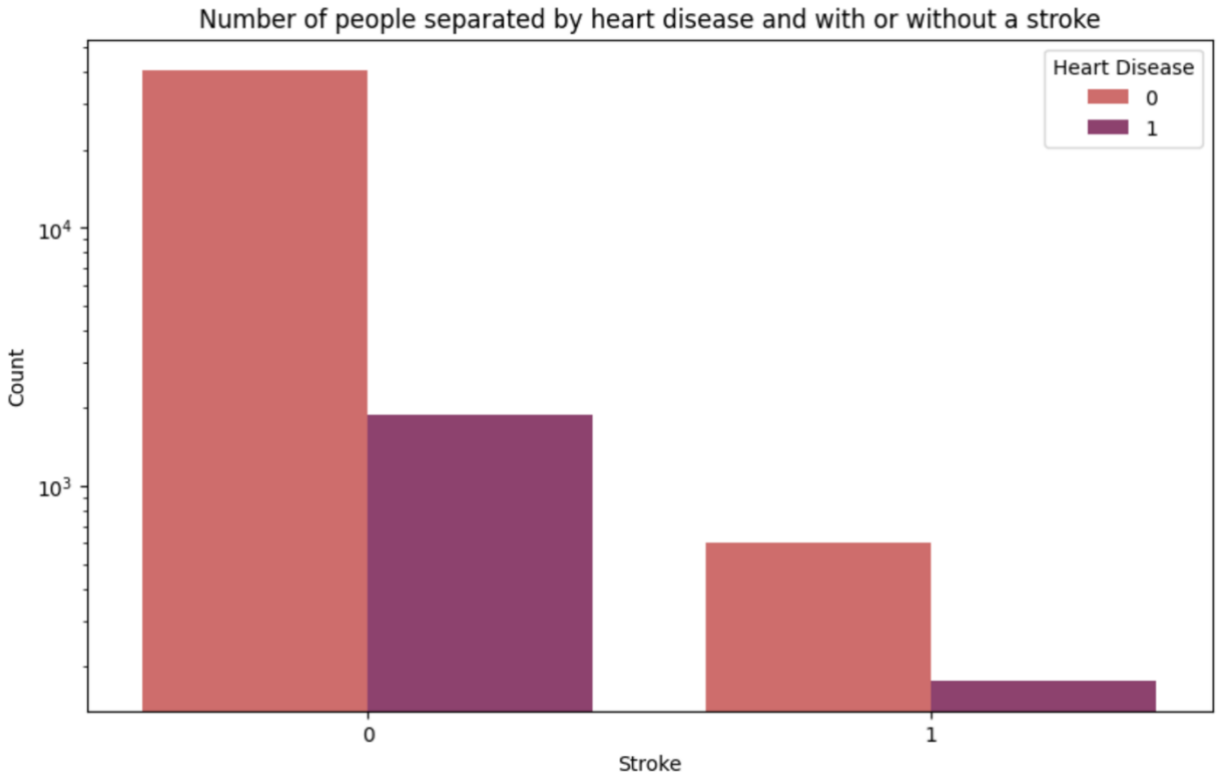

Figure 0.6: Heart disease distribution.

3.3.5 Marital status distribution

Figure 3.7 illustrates the distribution of individuals based on marital status and whether they have had a stroke. The x-axis differentiates between those who have not had a stroke (0) and those who have (1), while the y-axis measures the count of individuals. The bars are color-coded to reflect marital status, with one color representing individuals who have never been married and another for those who have been married at least once. The visualization suggests that among those who have and have not experienced a stroke, a higher proportion have been married. Proving that the marital status does not have effective impact on stroke occurrence.

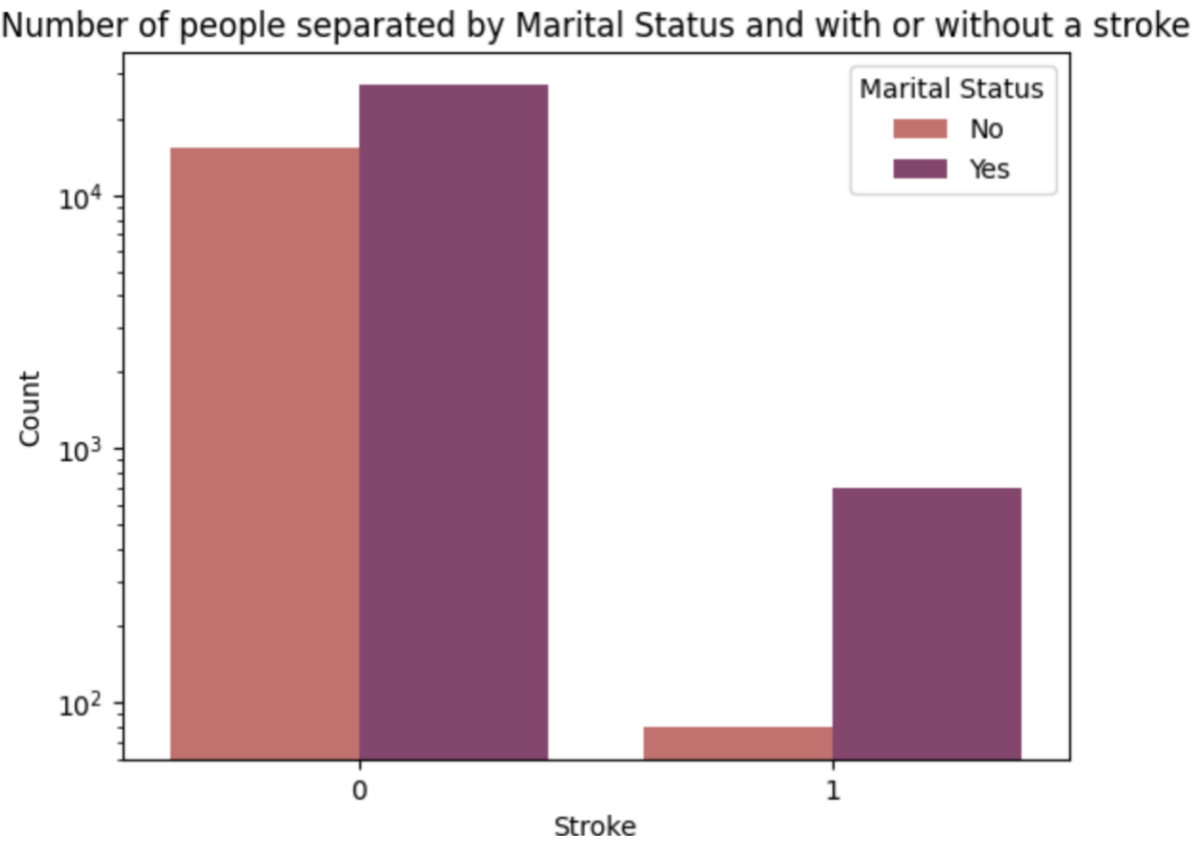

Figure 0.7: Martial status distribution

3.3.6 Work type distribution

Figure 3.8 presents a comparison of stroke incidence across different work types. The x-axis represents stroke status with '0' indicating individuals without a stroke and '1' indicating those who have had a stroke. The y-axis shows the count of individuals. The bars are color-coded to differentiate among work types: children, private sector employment, never worked, self-employment, and government jobs. This visualization allows for an analysis of how occupational factors may correlate with stroke occurrence, with private sector employment showing the highest count among those without a stroke. The data could suggest that certain work types are associated with either a higher or lower incidence of stroke.

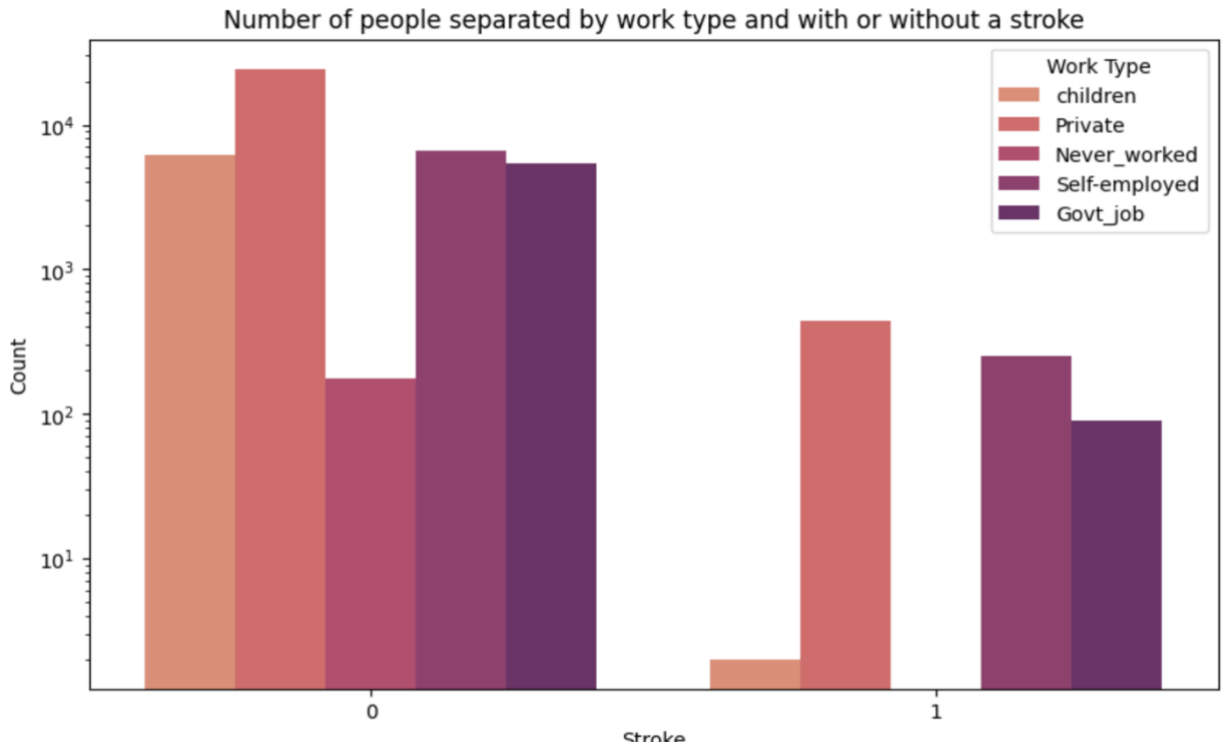

Figure 0.8: Work type distribution

3.3.7 Residence type distribution

Figure 3.9 delineates the number of individuals according to their residence type, further categorized by stroke occurrence. The x-axis indicates whether individuals have experienced a stroke, with '0' representing those who have not had a stroke and '1' for those who have. The y-axis represents the count of individuals. Bars are color-coded to differentiate between 'Rural' and 'Urban' residence types. The visualization suggests a comparison in the prevalence of stroke between rural and urban residents within the study's population.

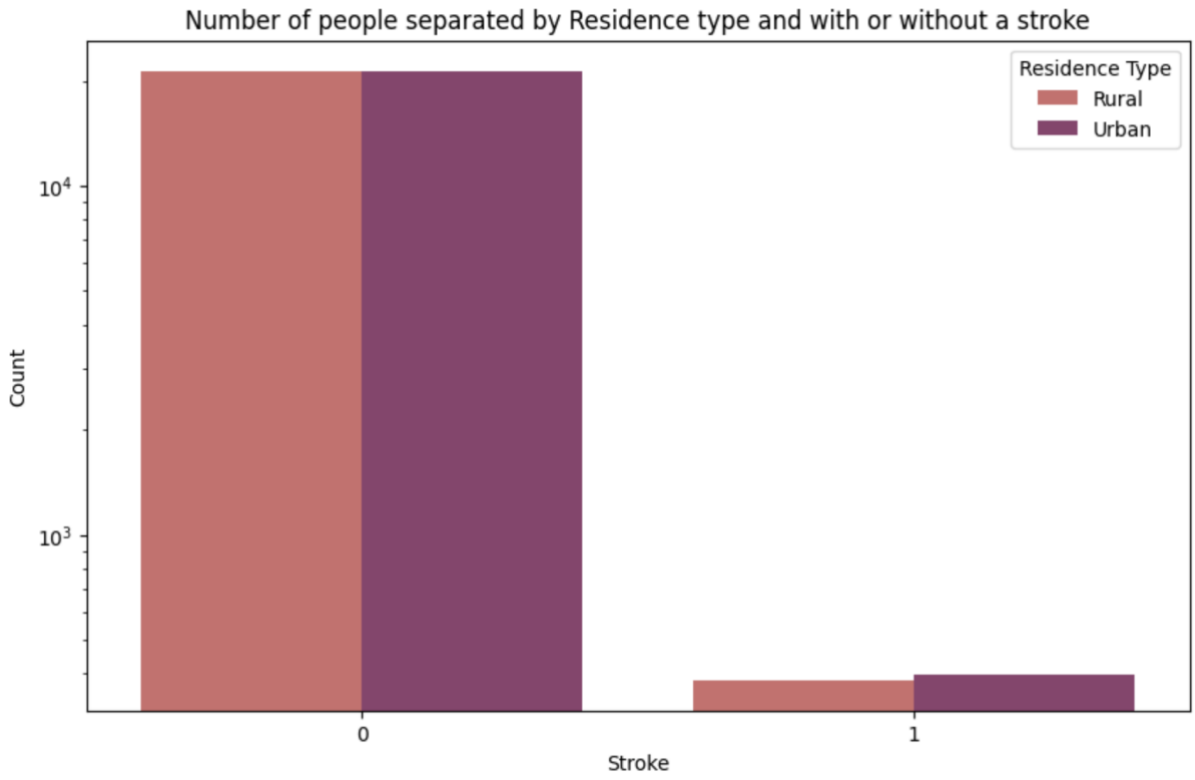

Figure 0.9: Residence type distribution.

Figure 3.10 illustrates the percentage distribution of stroke patients residing in rural and urban settings. It shows a nearly even split between the two residence types, with a slight majority living in rural areas. This visual comparison may indicate the proportional impact of geographical living conditions on stroke prevalence within the dataset.

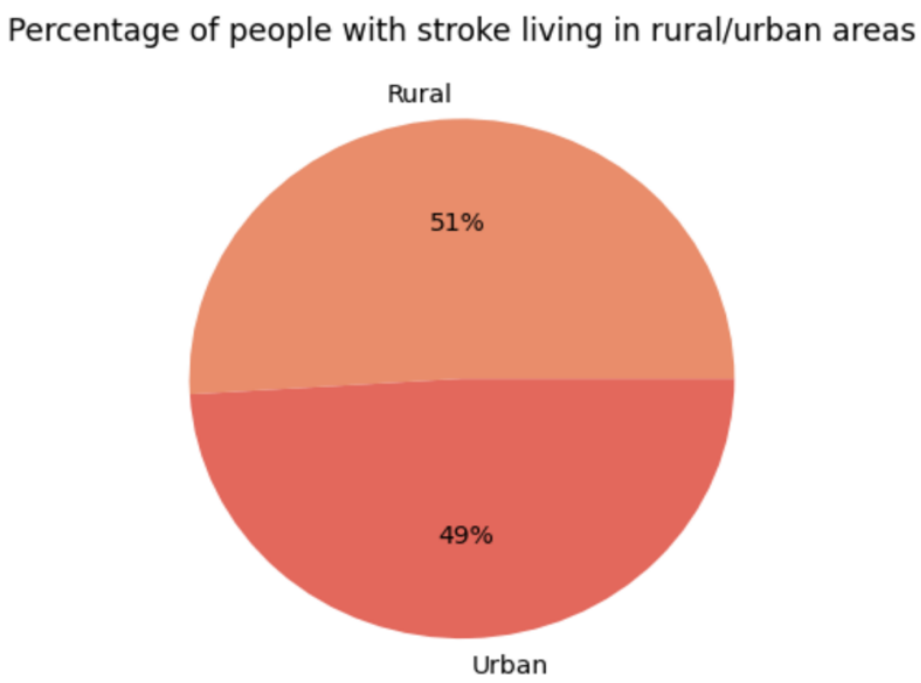

Figure 0.10: Proportional Distribution of Stroke Patients by Living Area: Rural versus Urban.

3.3.8 Average glucose distribution

Figure 3.11 illustrates the distribution of average glucose levels among individuals who have experienced a stroke. The x-axis represents the average glucose level, while the y-axis shows the count of individuals within each glucose level range. The bars depict the frequency of individuals in each glucose level category, with a superimposed density curve to highlight the overall distribution pattern. This visualization helps in understanding the spread and concentration of average glucose levels among stroke patients, indicating potential glucose level ranges that are associated with stroke occurrences.

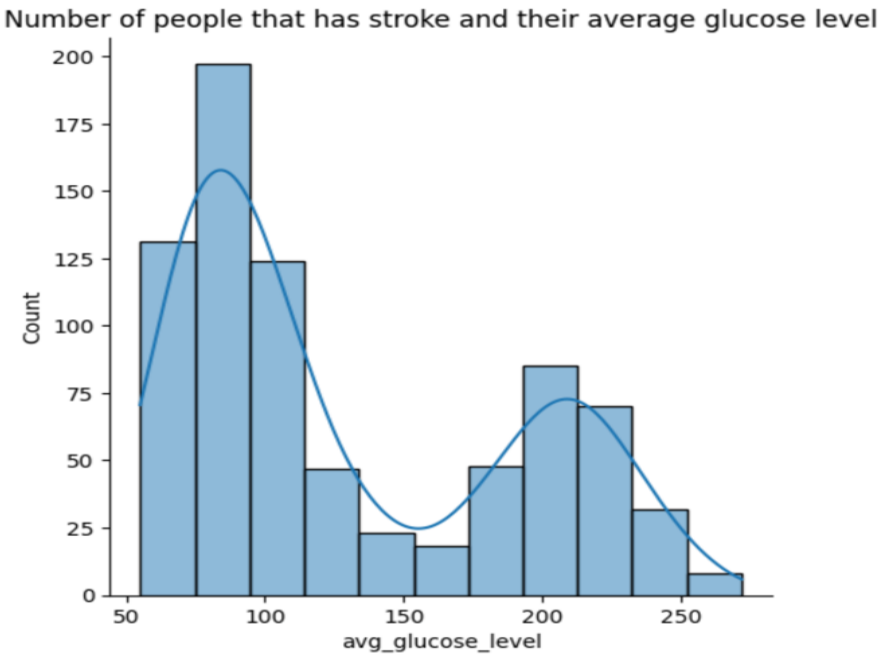

Figure 0.11: Distribution of Average Glucose Levels in Stroke Patients.

Figure 3.12 presents the distribution of average glucose levels in individuals with and without a stroke. The x-axis indicates the average glucose level, and the y-axis represents the density of individuals. The plot differentiates between individuals with a stroke (denoted by '1') and without a stroke (denoted by '0') using distinct colors. This chart provides a clear visualization of how average glucose levels vary between the two groups, offering insights into the relationship between glucose levels and stroke risk. By comparing the density curves, one can observe the differences in glucose level distributions, which may suggest a higher prevalence of certain glucose levels among stroke patients.

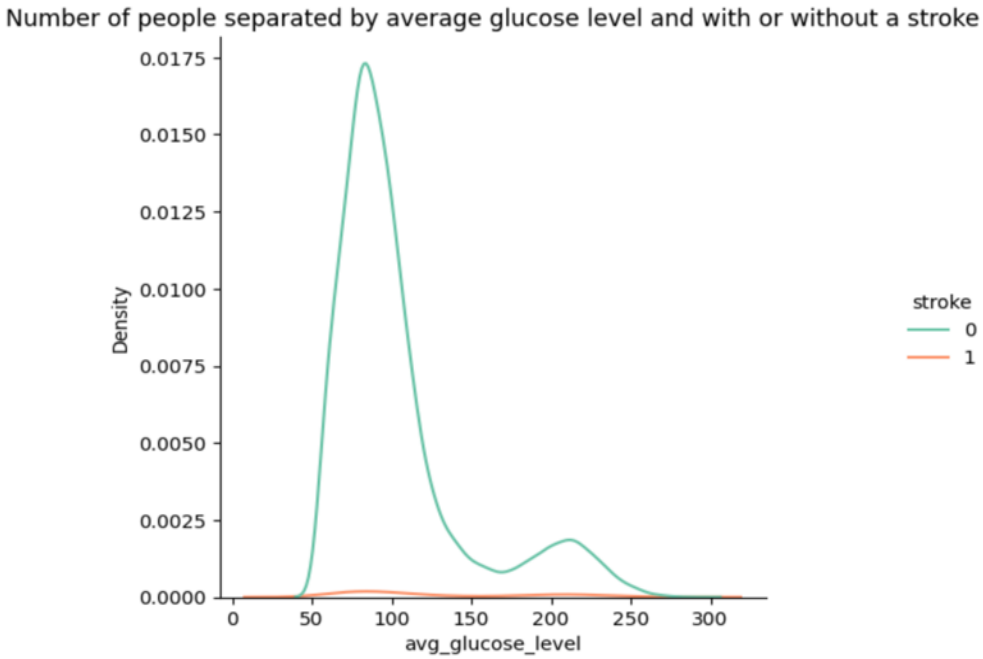

Figure 0.12: Average glucose distribution.

3.3.9 BMI distribution

Figure 3.13 presents the distribution of BMI among individuals who have experienced a stroke. The x-axis indicates BMI values, while the y-axis shows the count of individuals within each BMI range. The bars display the frequency of stroke patients in each BMI category, with a superimposed density curve to illustrate the overall distribution pattern. This chart helps in understanding the spread and concentration of BMI values among stroke patients, showing that most stroke occurrences are concentrated in the BMI range of 20 to 40. This distribution suggests that higher BMI may be a significant factor in the prevalence of stroke within the studied population.

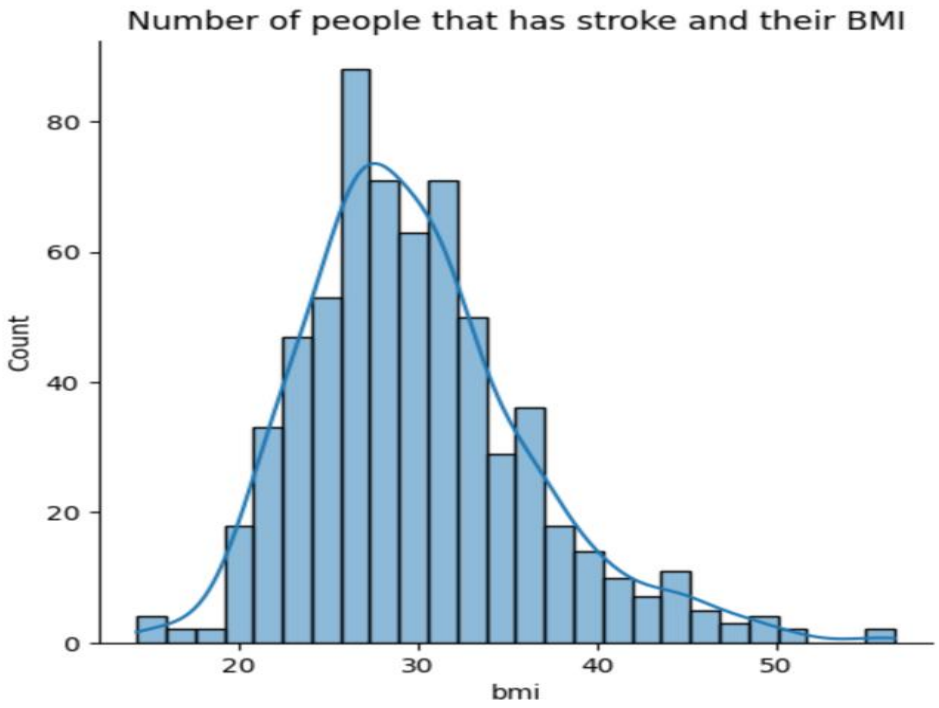

Figure 0.13: Distribution of BMI Among Individuals Who Have Suffered a Stroke.

Figure 3.14 shows the distribution of Body Mass Index (BMI) among individuals with and without a stroke. The x-axis represents BMI values, and the y-axis represents the density of individuals. The plot uses different colors to distinguish between individuals who have suffered a stroke (denoted by '1') and those who have not (denoted by '0'). This visualization provides insights into the variation in BMI between the two groups, indicating potential differences in BMI distributions that could be associated with stroke risk. The density curves highlight that a higher proportion of individuals with a stroke have higher BMI values compared to those without a stroke.

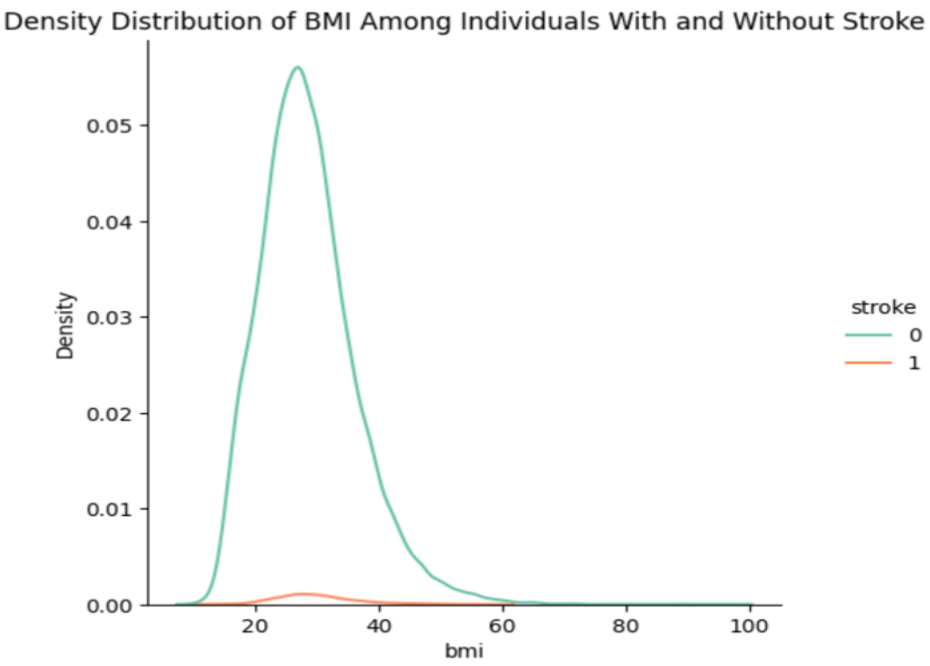

Figure 0.14: BMI distribution.

3.3.10 Smoking distribution

Figure 3.15 displays the smoking status of individuals who have experienced a stroke, segmented by gender. The x-axis categorizes smoking status into 'formerly smoked', 'never smoked', 'smokes', and 'Missing' for unreported data. The y-axis represents the count of individuals in each category. Bars are color-coded to differentiate between male and female genders. This visualization highlights the distribution of smoking habits among stroke patients and indicates the presence of missing data within the smoking status variable. It appears that among stroke patients, the 'never smoked' category has the highest count, particularly in females, followed by 'formerly smoked' and 'smokes', with a notable portion of missing data.

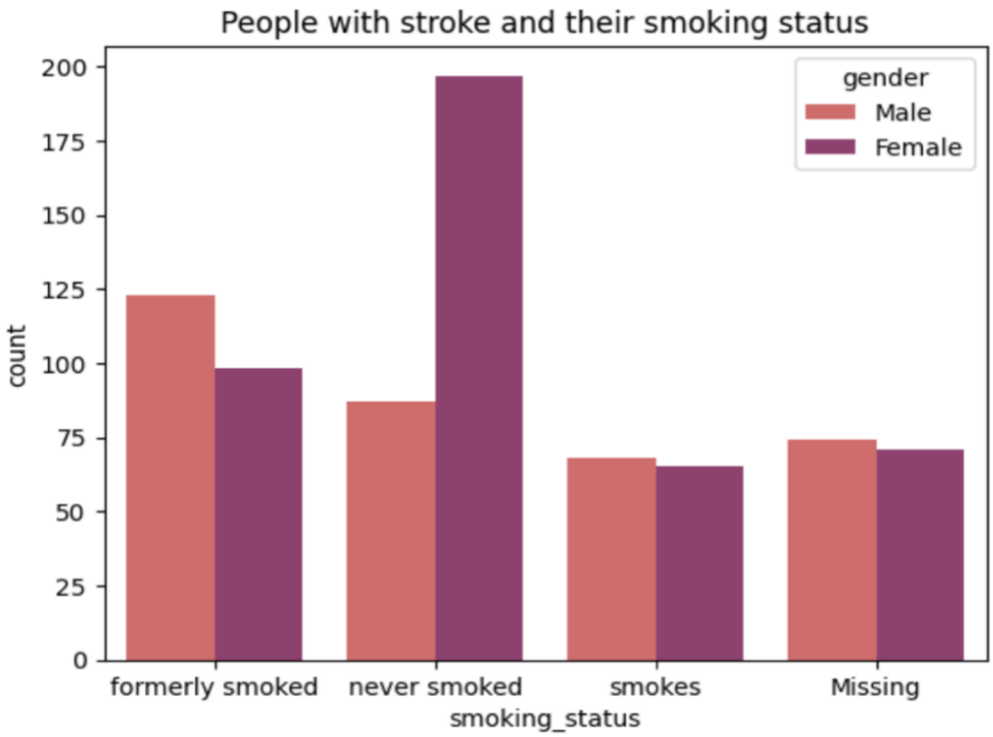

Figure 0.15: Smoking distribution.

3.4 Data preprocessing

Dataset preprocessing stands as a cornerstone in the application of machine learning and deep learning models, particularly in medical predictions like stroke occurrence. This initial phase is pivotal for refining the dataset, ensuring that subsequent analyses are based on accurate, consistent, and meaningful data. The preprocessing journey encompasses various techniques tailored to enhance data quality and model readiness, addressing common issues such as missing values, noise, and irrelevant features.

Preprocessing in the realm of ML and DL transcends basic data cleaning; it involves transforming the dataset into a format that machine learning and deep learning algorithms can efficiently process and learn from. This includes normalizing data ranges, encoding categorical variables into

numerical formats, Dataset balancing and handling missing data – a prevalent challenge in real-world datasets.

3.4.1 Missing values handling

Missing data can significantly skew predictions, leading to biased or inaccurate models. The approach to handling missing values must be judiciously selected, balancing between data integrity and the practicality of the solution. Common strategies include deletion, where incomplete records are removed, and imputation is used where missing values are replaced with substitute values based on various rationale.

Deletion of records

In the realm of data preprocessing for machine learning and deep learning, deletion stands as a straightforward yet decisive method for managing missing values. This approach, also known as listwise or case wise deletion, involves removing entire records from the dataset where any value is missing. While this method can be effective in datasets with a small percentage of missing values, it assumes that the missingness is completely at random, which may not always be the case. If data are not missing at random, deletion can introduce bias, leading to skewed results and potentially undermining the reliability of the model's predictions.

The concept of "Missing at Random" (MAR) plays a critical role in the methodology selected for data imputation and the subsequent reliability of statistical analyses and model predictions. MAR posits that while the propensity for data to be missing is not uniform across the dataset, it is conditional upon the observed data rather than the missing data itself. This implies that any pattern

in the missingness can be fully explained by variables for which data are available, without dependence on the values of the missing data.

Contrastingly, the "Missing Completely at Random" (MCAR) mechanism denotes an ideal scenario where the occurrence of missing data is entirely independent of both observed and unobserved data, rendering the missingness entirely stochastic. On the other end of the spectrum, "Missing Not at Random" (MNAR) describes situations where the probability of data being missing is directly influenced by the unobserved data, presenting a complex challenge for unbiased data imputation and analysis.

For this research, deletion was considered as a potential method for handling missing values. However, its applicability was rigorously evaluated against the dataset's characteristics and the requirements of the machine learning or deep learning model. Given that the integrity of the dataset is paramount and the volume of data sufficient, deletion should be employed only where it does not compromise the representativeness of the dataset or the systemic patterns within the data. This ensured that the resultant dataset remained robust and conducive to developing an accurate and generalizable machine learning and deep learning model for stroke prediction.

Under the gender feature, the category "Other" was encountered. Given the context of this study and the comparative analysis between male and female categories in relation to stroke prediction, it was deemed necessary to consolidate the "Other" classification. Considering the analytical framework's requirements, these 11 records were systematically excluded from the dataset. The exclusion of the "Other" category was a measured approach to streamline the analytical process, ensuring that the model's interpretive clarity and the research's overall findings remained uncompromised. This decision was driven by the need to simplify the gender variable for the

purposes of this specific analysis. This exclusion aimed to reduce complexity within the model without significantly impacting the overall findings of the research. It's important to note that this decision was specific to the constraints and objectives of this study.

Imputation

Imputation in data preprocessing is a crucial method for handling missing values, especially in the context of preparing datasets for machine learning and deep learning models. As a comprehensive approach, imputation addresses the absence of data by substituting missing entries with estimated values, thus allowing researchers to maintain the full breadth of collected data without resorting to the deletion of incomplete records.

The significance of imputation lies in its ability to preserve valuable data, especially in cases where the removal of records could result in the loss of critical information. This is particularly relevant in medical datasets, such as those used for stroke prediction, where each variable can be a vital predictor and every patient record is significant. By using imputation, we ensure that the patterns and relationships inherent in the complete cases are extended to fill the gaps in incomplete ones.

In the realm of data imputation, simple techniques such as mean, median, and mode substitution are commonly employed to address the issue of missing values. Each method has distinct characteristics that render it suitable for different scenarios, largely influenced by the distribution of the data in question.

Multivariate imputation is more sophisticated approach, which considers the relationships between multiple variables to estimate the missing values. One such method is Multivariate

Imputation by Chained Equations (MICE) [42][43], which performs multiple regressions over the dataset and draws values from the predictive distribution to fill in the missing data iteratively.

For the current thesis, the chosen method of imputation was multivariate imputation using the Iterative Imputer with Bayesian Ridge as the estimator. This method was applied to handle missing values for the BMI and smoking status features. Multivariate imputation was selected because it considers the relationships among multiple variables in the dataset, providing a more accurate and consistent approach to imputing missing data.

The decision to use multivariate imputation was supported by an analysis of the missingness pattern within the dataset. It was observed that the missing values did not occur systematically and were spread randomly across records. This non-systematic nature of missingness indicated that a multivariate approach would be suitable for capturing the underlying patterns in the data and ensuring that the imputed values were representative of the actual distributions.

Under the smoking status feature, instances labeled as "Unknown" presented a challenge. The presence of "Unknown" values could introduce uncertainty into the model, potentially skewing the analysis and leading to less accurate predictions. To address this issue, a decision was made to treat "Unknown" values as missing values to be imputed using multivariate imputation. This choice helped to maintain the integrity of the dataset by ensuring that all entries under the smoking status feature contributed meaningfully to the model's learning process.

The replacements under the smoking status features were implemented as part of the data normalization and categorical value consolidation process. By making these adjustments, the dataset was better positioned for the subsequent steps of categorical variable encoding. These preprocessing decisions were critical for preparing the dataset, ensuring that each feature was

encoded effectively and contributed positively to the predictive model's ability to learn from the data.

In conclusion, imputation serves as a robust strategy to counter the challenges posed by missing data. The appropriate use of multivariate imputation within this thesis underscores its value in maintaining the quality and integrity of the dataset, which is paramount for the development of reliable machine learning and deep learning models for medical predictions.

3.4.2 Dataset Shuffling

Dataset shuffling is a crucial step in data preprocessing for machine learning and deep learning models, particularly in ensuring the integrity and generalizability of predictive models. Shuffling the dataset involves randomly reordering the rows of data, which helps in mitigating any inherent biases that could affect the training and evaluation processes.

The primary purpose of shuffling the dataset is to ensure that the data fed into the machine learning or deep learning model is randomly distributed. This randomness is essential for creating robust models that generalize well to new, unseen data. Without shuffling, the model might learn patterns that are not representative of the entire dataset, especially if the data has any underlying order that could influence the learning process.

In datasets with time-series or sequential data, where the order of data points is significant, shuffling is typically not recommended. However, for most other types of data, especially those used in classification tasks like stroke prediction, shuffling helps in breaking any potential correlations between consecutive data points.

Shuffling the dataset has a significant impact on the performance of the machine learning and deep learning model. By ensuring that the training set is diverse and representative of the entire dataset, the model can learn more generalizable patterns. This, in turn, leads to improved performance on the testing set and better predictive capabilities when applied to new data.

3.4.3 Categorical Variable Encoding

Categorical Variable Encoding is an indispensable step in data preprocessing for machine learning and deep learning models. It involves transforming categorical data, which is often text-based, into a numerical format that can be understood and processed by machine learning and deep learning algorithms. This transformation is critical because while human cognition can easily interpret and classify text data, machine learning and deep learning models require numerical input to perform tasks such as classification, regression, or prediction.

Understanding Categorical Data

Categorical data comes in two forms: nominal and ordinal. Nominal data represents discrete units without an inherent order, such as color or zip codes, while ordinal data maintains a natural ranking, like educational level or job grades. Properly encoding this data ensures that the machine learning and deep learning algorithms capture the correct relationships present in the data structure.

Label Encoding

This technique assigns a unique integer to each category. Label encoding [44] is particularly beneficial for ordinal data since the assigned numbers can often represent a hierarchy in the data. For instance, in a feature representing education levels, 'high school' could be encoded with a lower

number than 'bachelor's degree,' which in turn could be lower than 'master's degree,' preserving the educational hierarchy. The simplicity and computational efficiency of Label Encoding make it a suitable choice for datasets with ordinal categories. However, this method can introduce a potential problem when applied to nominal data: it implies an artificial order that may not exist, which could lead to misinterpretation by the model.

Validation of Encoding choices: The encoded dataset was meticulously scrutinized to validate the efficacy of the encoding techniques employed. By training machine learning or deep learning models using the transformed dataset and evaluating them against a suite of performance metrics, including accuracy, Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), and F1-score, the encoding's validity was corroborated. Comparative analysis with models trained on data processed with alternative encoding methods, as well as the raw dataset, provided a benchmark for assessing the retention of the categorical variables' predictive capacity. This thorough approach confirmed that the chosen encoding scheme is not only preserved but in some cases, augmented the informational essence of the categorical features, ensuring their substantial contribution to the predictive process was maintained.

Categorical Variable Encoding stands as a critical process in the data preparation stage for machine learning and deep learning models. The careful application of Label Encoding in this thesis has demonstrated that with a judicious choice of encoding method, one can maintain the integrity of the data while transforming it into a format conducive to model training and subsequent stroke prediction. This step has facilitated the use of advanced machine learning and deep learning techniques, which hinge on the numerical representation of data, to uncover patterns and insights that are pivotal in predicting stroke occurrences.

3.4.4 Imbalanced class handling

Handling imbalanced classes is a common and critical issue in machine learning and deep learning, particularly in domains like medical diagnostics where the event of interest (such as a stroke) may be rare compared to the non-event cases. An imbalanced dataset can bias the model towards the majority class, resulting in poor performance when predicting the minority class. Therefore, techniques to handle class imbalance are essential to ensure the robustness and accuracy of predictive models.

Oversampling Techniques

Oversampling techniques address class imbalance by augmenting the minority class to match the prevalence of the majority class. This strategy enhances the dataset's balance, aiming to improve model performance on underrepresented classes.

Adaptive Synthetic Sampling (ADASYN): ADASYN [43] is an oversampling technique that generates synthetic samples of the minority class based on the data distribution. It operates by creating more synthetic data for minority class instances that are harder to learn, as determined by the number of majority class instances in their neighborhood. This results in a more balanced dataset, which in turn can lead to a more robust and fair learning process.

The ADASYN algorithm is particularly well-suited for datasets where the imbalance is not only prevalent but also where the decision boundary between classes is not well defined. By generating synthetic samples in regions of the feature space where the minority class is underrepresented, ADASYN helps in constructing a more complex decision function that can capture the nuances and patterns specific to the minority class.

Implementing ADASYN involved several steps:

1. **Analyzing the Dataset**: The initial step was to analyze the degree of imbalance in the dataset by calculating the ratio of the number of instances in the minority class to the majority class.

2. **Applying ADASYN**: After determining the imbalance ratio, ADASYN was applied to the dataset to synthesize new, artificial minority class instances. This was achieved using the "**imblearn**" library in Python.

3. **Training and Evaluation**: The resampled dataset was used to train the machine learning and deep learning models. The model's performance was evaluated using metrics that are sensitive to class imbalance, such as the F1 score, and the area under the Receiver Operating Characteristic (ROC) curve.

The application of ADASYN proved to be beneficial for the stroke prediction model. There was a noticeable improvement in the prediction accuracy for the minority class without a significant loss in the majority class's prediction accuracy. This indicated that the model became more general and capable of identifying patterns across both classes.

The use of ADASYN also introduced some considerations and potential limitations to be addressed. The synthetic samples generated by ADASYN can sometimes lead to overfitting if the algorithm over-emphasizes the minority class regions that are already well-represented. To mitigate this, a careful tuning of ADASYN's parameters was necessary, particularly the "**n_neighbors"** parameter, which controls the adaptive nature of the synthetic sample generation.

Tuning the "n_neighbors" Parameter: The parameter "n_neighbors" plays a central role in ADASYN's operation. It determines the number of nearest neighbors used to form the synthetic samples. Tuning this parameter involves a delicate balance:

1. **Lower Values of "n_neighbors"**: Setting a lower value may concentrate synthetic data generation around existing minority examples, risking over-representation of those areas.

2. **Higher Values of "n_neighbors"**: Conversely, higher values could lead to broader synthetic data generation but may dilute the adaptive nature intended to focus on difficult-to-classify regions.

Careful tuning implies a systematic, data-driven approach to adjusting "n_neighbors". Employing a strategy such as a grid search allows for the comprehensive exploration of "n_neighbors" values within a predefined range. This exploration is instrumental in understanding the parameter's influence on key performance metrics of the learning algorithm—namely, accuracy, precision, recall, and the F1 score. By scrutinizing the outcomes of these metrics, one can discern the optimal "n_neighbors" setting that harmonizes with the algorithm’s predictive capabilities, thus enhancing the model's efficacy in distinguishing between classes.

The effectiveness of ADASYN was also compared against other oversampling techniques, such as Random Undersampling, and Random Oversampling. The comparison was based on the overall performance of the trained model. ADASYN was found to achieve a higher model performance metrics.

In conclusion, handling imbalanced classes is a crucial step in preparing datasets for machine learning and deep learning models in medical diagnostics. The adoption of ADASYN for class

balancing in the stroke prediction dataset facilitated the development of a more accurate and reliable predictive model. This approach has shown that with careful application and parameter tuning, oversampling techniques like ADASYN can significantly enhance model performance, especially in the context of imbalanced medical data, thereby contributing to the advancement of predictive analytics in healthcare.

3.4.5 Data Normalization

Data normalization is a critical preprocessing step in machine learning and deep learning, particularly for algorithms that are sensitive to the scale of the data, such as machine learning models. This process involves transforming the data so that it fits within a specific scale, like 0–1 or a standard normal distribution with a mean of 0 and a standard deviation of 1. The primary goal of data scaling is to ensure that all features contribute equally to the result and to prevent features with larger ranges from dominating the model's learning.

For machine learning and deep learning models, which often involve gradient descent optimization, scaling helps in speeding up the convergence by providing a level playing field. Without scaling, the optimization process can become biased towards the features with larger magnitudes, leading to an inefficient learning process and a model that may not generalize well.

Min-Max Scaling: Min-Max Scaling method [45] rescales the data to a fixed range, typically [0, 1]. The min-max scaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum.

Chapter 4

Methods

After completing exploratory data analysis (EDA), handling missing data, data imputation, encoding categorical variables, data normalization, handling imbalanced dataset, the next step is model development and training. This involves selecting appropriate machine learning or deep learning models, training these models on the prepared datasets, and then evaluating their performance.

4.1 Classification Methods

Classification in machine learning and deep learning is a supervised learning task where the goal is to categorize input data into predefined classes or labels based on its features. During the training phase, the algorithm is provided with a labeled dataset, where each data point has a set of features and an associated class label. The primary objective is to learn a mapping from these features to the correct class labels, enabling the algorithm to generalize its learning and accurately classify new, unseen instances.

The classification process involves identifying patterns and relationships within the training data to create a model that can make predictions on new data. Various algorithms can be used for classification tasks, each with its own strengths and methodologies. The effectiveness of classification methods lies in their ability to learn efficiently from labeled data, generalize patterns, and make accurate predictions on new instances. This capability is crucial for advancing automated decision-making across various domains.

Different classification algorithms have unique characteristics and applications. In this study, key algorithms including Gradient Boosting Classifier, Extreme Gradient Boosting, and Adaboost among others were implemented.

4.1.1 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that extends the basic concept of decision trees to create a more robust and accurate predictive model. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This approach helps mitigate the overfitting problem commonly associated with single decision trees [46].

In Random Forest, each tree in the ensemble is built from a bootstrap sample of the training data. Additionally, when splitting nodes, the algorithm considers a random subset of features, rather than all features. This randomization helps in producing diverse trees whose combined output is less prone to overfitting [47]. The combined effect of bootstrapping and random feature selection leads to a reduction in variance and improved model performance.

One of the key advantages of Random Forest is its ability to handle large datasets with higher dimensionality effectively. It is robust to outliers and can capture complex interactions among features. Moreover, Random Forest provides an intrinsic measure of feature importance, which helps in understanding the contribution of different features to the model's predictions [48].

However, Random Forests can be computationally intensive and require more memory due to the creation of multiple trees. Despite this, they are highly parallelizable, which can mitigate computational costs when using modern multi-core processors.

In the context of stroke prediction, Random Forest can be utilized to classify patients based on various clinical and demographic features. The algorithm's ensemble nature and ability to handle mixed data types make it particularly suitable for this task. By analyzing the ensemble of decision trees, Random Forest can provide insights into which features are most critical for predicting stroke risk, thus supporting early diagnosis and personalized treatment planning.

4.1.2 Decision Tree (DT)

Decision Trees (DT) are a popular and intuitive classification algorithm widely used in machine learning. They operate by recursively splitting the dataset into subsets based on the value of the input features, creating a tree-like model of decisions. Each node in the tree represents a feature in the data, and each branch represents a decision rule, leading to a final decision or class label at the leaf nodes [50].

The process of building a decision tree involves selecting the feature that best splits the data into distinct classes at each step. This selection is typically based on criteria such as Gini impurity or information gain, which measure the quality of the split. The goal is to create a tree that accurately classifies the training data while maintaining generalizability to unseen data [49].

One of the main advantages of decision trees is their interpretability. The tree structure is easy to visualize and understand, allowing practitioners to trace the path from input features to the final decision. This makes decision trees particularly valuable in fields where model interpretability is crucial, such as healthcare and finance. Additionally, decision trees can handle both numerical and categorical data and are robust to outliers [51].

However, decision trees are prone to overfitting, especially when they grow too deep and capture noise in the training data. Pruning techniques, such as cost-complexity pruning, are often employed to mitigate this issue by removing branches that have little importance in the overall model [49].

In the context of stroke prediction, decision trees can be used to classify patients based on various clinical and demographic features. The algorithm's ability to model complex decision boundaries and its interpretability make it a valuable tool for predicting stroke risk. By analyzing the decision paths, clinicians can gain insights into which features are most influential in determining stroke risk, aiding in early diagnosis and personalized treatment planning.

4.1.3 Gradient Boosting Classifier (G-Boost)

Gradient Boosting Classifier, often referred to as G-Boost, is a powerful machine learning algorithm designed explicitly for classification tasks. It operates within the ensemble learning framework, combining the outputs of multiple weak learners, typically decision trees, in a sequential manner. This approach allows the model to iteratively correct errors and enhance its predictive accuracy, making it highly effective at capturing complex patterns within the data [52].

The core of the Gradient Boosting Classifier lies in its optimization through gradient descent. By iteratively fitting decision trees to the model's residuals, the algorithm refines its predictions, providing a sophisticated mechanism to navigate intricate relationships in the data. The use of weak learners, often shallow decision trees, contributes to the overall strength of the ensemble, compensating for individual limitations and resulting in a robust model [53].

A notable feature of G-Boost is its incorporation of regularization techniques, which help prevent overfitting by introducing penalty terms during optimization. This ensures that the model not only

captures intricate patterns in the training data but also generalizes effectively to unseen data. Tree pruning is also implemented to eliminate less influential branches, optimizing the structure of each decision tree for improved efficiency and interpretability [52].

The algorithm's scalability is commendable, owing to its parallel processing capabilities and compatibility with distributed computing, making it particularly efficient for large datasets in real-world applications. Moreover, G-Boost provides valuable insights into feature importance, transparently revealing the variables that significantly contribute to the model's predictions [53].

In the context of stroke prediction, G-Boost is particularly valuable due to its ability to handle various features. By sequentially adjusting for the errors of previous iterations, the model can capture the subtle interactions between variables that are indicative of stroke risk. This capability makes it particularly suited for complex medical datasets where multiple factors contribute to the outcome. The robustness and high predictive accuracy of the Gradient Boosting Classifier make it a valuable tool in the early detection and prevention of strokes.

4.1.4 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting, commonly known as XGBoost, is an advanced machine learning algorithm that belongs to the ensemble learning family. It is a gradient-boosting algorithm that combines the strengths of decision trees with regularization techniques to create a highly potent and precise predictive model. XGBoost is widely recognized for its superior performance in both regression and classification tasks, making it popular across diverse domains [70]

XGBoost operates within the gradient boosting framework, constructing an ensemble of decision trees sequentially. Each iteration focuses on refining the predictions by addressing the errors made

by the previous models. This method allows XGBoost to effectively manage complex relationships within the data, enhancing predictive accuracy [55] One of the standout features of XGBoost is its incorporation of regularization techniques, which help prevent overfitting by introducing penalty terms during optimization. This ensures that the model captures intricate patterns in the training data while maintaining good generalization to unseen data [54].

Additionally, XGBoost implements tree pruning to remove less influential branches, optimizing the structure of each decision tree for improved efficiency and interpretability. The algorithm's scalability is enhanced by its parallel processing capabilities and compatibility with distributed computing, making it particularly efficient for large datasets in real-world applications [54].

XGBoost also provides valuable insights into feature importance, transparently revealing the variables that significantly contribute to the model's predictions. This transparency is especially useful for understanding which features are most influential in making predictions [55].

The versatility and robustness of XGBoost have made it a fundamental tool for data scientists and practitioners in fields such as finance, healthcare, and natural language processing. Its ability to balance model complexity, interpretability, and computational efficiency makes it a preferred choice for developing high-performance predictive models in many practical applications [54].

In the context of stroke prediction, XGBoost is particularly valuable due to its ability to integrate regularization techniques, crucial for handling clinical datasets prone to overfitting. By sequentially constructing decision tree ensembles within the gradient boosting framework, XGBoost can navigate the intricate relationships within clinical data, continuously enhancing its predictive accuracy for stroke risk. This iterative process allows the model to adeptly manage the

complexities of clinical variables, thereby improving its ability to predict the likelihood of stroke occurrence.

4.1.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, yet effective, non-parametric classification algorithm widely used in machine learning. It operates on the principle of similarity, where the classification of a new instance is determined based on the majority class of its k-nearest neighbors in the feature space. This method does not make any underlying assumptions about the data distribution, making it versatile and applicable to various types of data [56].

In the KNN algorithm, the value of k, which denotes the number of nearest neighbors to consider, is a critical parameter. Selecting an appropriate k is essential for the performance of the model. A small k value can lead to a noisy decision boundary and overfitting, whereas a large k value might smooth out the decision boundary too much, causing underfitting. Typically, k is chosen through cross-validation to optimize model performance [57].

The algorithm works by calculating the distance between the new instance and all the instances in the training dataset. Common distance metrics include Euclidean, Manhattan, and Minkowski distances. The k instances with the smallest distances are identified, and the majority class among these neighbors determines the class label of the new instance [56].

KNN is particularly useful for its simplicity and ease of implementation. It is also highly interpretable, as the decision for classifying a new instance is directly influenced by the closest training examples. However, KNN can be computationally expensive, especially with large datasets, since it requires computing the distance to every training instance for each prediction.

Additionally, KNN is sensitive to the scale of the data, so normalization or standardization of features is often necessary.

In the context of stroke prediction, KNN can be applied to classify patients based on various clinical and demographic features. The algorithm's ability to consider multiple features and find similar cases in the training data makes it a valuable tool for predicting stroke risk. By examining the nearest neighbors, KNN can identify patterns and similarities that are indicative of stroke, thus aiding in early diagnosis and intervention.

4.1.6 Logistic Regression (LR)

Logistic Regression (LR) is a widely used statistical method for binary classification tasks, where the outcome variable has two possible categories. Despite its name, logistic regression is primarily used for classification rather than regression. This algorithm is particularly effective in predicting the probability of an event occurring based on one or more predictor variables [58].

In logistic regression, the logistic (or sigmoid) function transforms a linear combination of input features into a probability value between 0 and 1. If this probability exceeds a certain threshold (commonly 0.5), the model predicts the positive class; otherwise, it predicts the negative class. The logistic function ensures that the output is bounded between 0 and 1, making it suitable for binary classification tasks [59].

Logistic regression is well-suited for problems where the dependent variable is categorical, making it widely applicable in fields such as medicine (e.g., disease diagnosis), finance (e.g., credit scoring), and social sciences (e.g., voter prediction). It is a straightforward yet powerful algorithm

that provides interpretable results, allowing practitioners to understand the impact of individual predictors on the likelihood of a specific outcome [58].

One significant advantage of logistic regression is its interpretability. The coefficients assigned to each variable in the model indicate the direction and strength of their influence on the prediction. This interpretability is crucial in fields where understanding the underlying factors contributing to predictions is essential, such as healthcare [59].

In the context of stroke prediction, logistic regression can be applied to both numerical and categorical clinical variables, as well as patient demographics. The algorithm calculates the probability of a patient having a stroke. This probability is then compared to a predefined threshold, and if it exceeds the threshold, the model predicts the occurrence of a stroke; otherwise, it predicts no stroke. Logistic regression's simplicity, interpretability, and effectiveness make it a valuable tool for early diagnosis and intervention planning in stroke prediction.

4.1.7 Gaussian Naive Bayes (Gaussian-NB)

Gaussian Naive Bayes (Gaussian-NB) is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features. It is particularly suited for continuous data, where the likelihood of the features is assumed to follow a Gaussian (normal) distribution. This makes it a popular choice for various practical applications due to its simplicity and effectiveness [61].

The core principle of Gaussian Naive Bayes involves calculating the posterior probability of each class given the input features, using Bayes' theorem. The algorithm then selects the class with the

highest posterior probability as the predicted class. The Gaussian assumption simplifies the computation of these probabilities, making the algorithm computationally efficient [60].

One of the key advantages of Gaussian Naive Bayes is its simplicity. The algorithm requires a relatively small amount of training data to estimate the necessary parameters (means and variances of the features for each class). This makes it particularly effective for high-dimensional datasets where the curse of dimensionality might pose a problem for other algorithms [61].

Despite its simplicity, Gaussian Naive Bayes can perform surprisingly well in many real-world applications. However, its primary limitation lies in the strong independence assumption between features. In practice, this assumption is often violated, which can lead to suboptimal performance. Nonetheless, Gaussian Naive Bayes remains a robust baseline method and is often used for initial exploratory data analysis [60].

In the context of stroke prediction, Gaussian Naive Bayes can be used to classify patients based on various clinical and demographic features. By modeling the distribution of each feature as a Gaussian, the algorithm can effectively handle continuous variables such as age, blood pressure, and cholesterol levels. This probabilistic approach allows for straightforward interpretation of the model's predictions, aiding in early diagnosis and treatment planning for stroke patients.

4.1.8 AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm designed to improve the performance of weak classifiers. It combines multiple weak learners, typically decision trees with a single split (decision stumps), to create a strong classifier. AdaBoost works by iteratively

adjusting the weights of misclassified instances, focusing more on the difficult cases in subsequent rounds of training [62].

The core principle of AdaBoost is to iteratively train a sequence of weak learners, where each learner tries to correct the errors of its predecessor. After each round of training, the algorithm assigns higher weights to the misclassified instances and lower weights to correctly classified ones. This process continues until a predefined number of weak learners are combined or the model achieves a desired level of accuracy [62].

One of the key advantages of AdaBoost is its ability to improve the accuracy of weak classifiers significantly. It is particularly effective in reducing bias and variance, making it robust against overfitting. AdaBoost is also versatile, as it can be combined with various types of weak learners, not just decision stumps [63].

However, AdaBoost has some limitations. It is sensitive to noisy data and outliers because the algorithm assigns higher weights to misclassified instances, which can amplify the impact of noisy data. Despite this, AdaBoost remains a popular choice for many classification tasks due to its simplicity and effectiveness [63].

In the context of stroke prediction, AdaBoost can be applied to classify patients based on a combination of clinical and demographic features. By iteratively focusing on the harder-to-classify cases, AdaBoost enhances the overall model accuracy. This iterative boosting process allows the model to capture complex patterns in the data, thereby improving the prediction of stroke risk and aiding in early diagnosis and intervention.

4.1.9 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for both classification and regression tasks. It is particularly effective in high-dimensional spaces and is known for its robustness in handling non-linear data through the use of kernel functions [65].

The primary objective of SVM is to find the optimal hyperplane that maximizes the margin between the classes. This margin is defined as the distance between the hyperplane and the nearest data points from each class, known as support vectors. By maximizing this margin, SVM aims to improve the model's generalization ability [64].

SVM can handle non-linear classification problems using kernel functions such as polynomial, radial basis function (RBF), and sigmoid. These kernels transform the input features into a higher-dimensional space where a linear separator can be found. This flexibility allows SVM to model complex decision boundaries effectively [65].

One of the key advantages of SVM is its effectiveness in high-dimensional spaces, making it suitable for applications involving a large number of features. Additionally, SVM is relatively memory efficient because it uses a subset of training points (support vectors) in the decision function. However, SVM can be computationally intensive, especially with large datasets, and the choice of kernel and hyperparameters can significantly impact its performance [66].

In the context of stroke prediction, SVM can be used to classify patients based on clinical and demographic features. The algorithm's ability to create complex decision boundaries makes it well-suited for identifying patterns that are indicative of stroke risk. By selecting the appropriate kernel

and tuning the hyperparameters, SVM can provide accurate and reliable predictions, aiding in early diagnosis and personalized treatment planning.

4.1.10 Deep Neural Networks (DNN)

Deep Neural Networks (DNN) are a class of machine learning algorithms modeled after the human brain's neural networks. They consist of multiple layers of interconnected neurons, where each layer transforms the input data through learned weights and biases. DNNs are particularly powerful for capturing complex patterns and representations in data, making them suitable for a wide range of tasks including classification, regression, and more [68].

A typical DNN architecture comprises an input layer, several hidden layers, and an output layer. Each neuron in a layer receives inputs from neurons in the previous layer, applies a linear transformation followed by a non-linear activation function, and passes the result to the next layer. Common activation functions include ReLU (Rectified Linear Unit), sigmoid, and tanh, each contributing to the network's ability to model complex relationships [67]

Training a DNN involves optimizing the weights and biases using backpropagation and gradient descent. During backpropagation, the algorithm calculates the gradient of the loss function with respect to each weight by applying the chain rule, and then updates the weights in the opposite direction of the gradient to minimize the loss. This iterative process continues until the model converges to a solution that minimizes the error on the training data [68].

One of the key strengths of DNNs is their ability to automatically learn hierarchical feature representations from raw data. This capability makes them particularly effective for tasks involving high-dimensional data, such as image and speech recognition. However, training DNNs

can be computationally intensive and requires large amounts of labeled data to achieve high performance [67].

In the context of stroke prediction, DNNs can be used to classify patients based on a variety of clinical and demographic features. By leveraging the network's deep architecture, DNNs can capture intricate patterns and interactions within the data that may be indicative of stroke risk. This enables the development of accurate predictive models that can aid in early diagnosis and personalized treatment planning for stroke patients.

In this study, the DNN architecture used for stroke prediction consists of the following layers:

- **Input Layer:** This layer has neurons equal to the number of features in the dataset.
- **Hidden Layers:** The network includes three hidden layers:
    - The first hidden layer contains 256 neurons with ReLU activation.
    - The second hidden layer contains 256 neurons with ReLU activation.
    - The third hidden layer contains 128 neurons with ReLU activation and a dropout rate of 0.1.
    - The fourth hidden layer contains 128 neurons with ReLU activation and a dropout rate of 0.1.
    - The fifth hidden layer contains 32 neurons with ReLU activation and a dropout rate of 0.1.

- **Output Layer:** This layer has one neuron with a sigmoid activation function to predict the probability of stroke occurrence.

Dropout layers are included after certain hidden layers to reduce overfitting by randomly setting a fraction of input units to zero during training. This regularization technique helps improve the model's generalization ability.

Chapter 5

Results and Discussion

This chapter presents the findings from applying various machine learning and deep learning models to predict stroke occurrence. The models were evaluated using multiple performance metrics, including accuracy, precision, recall, F1 score, specificity, and the area under the receiver operating characteristic curve (AUC). The discussion provides a detailed comparison of the models' performance.

5.1 Evaluation Metrics

In the context of stroke prediction, evaluating the model's performance is crucial to understanding its effectiveness and reliability. The primary evaluation metrics used in this study are Accuracy, Area Under the Receiver Operating Characteristics (ROC) Curve (AUC), F1 Score, Precision, Recall, and Specificity. These metrics provide a comprehensive overview of the model's predictive capabilities and its ability to handle imbalanced datasets.

In binary classification for stroke detection, the model's predictions are categorized into four outcomes:

- **True Positives (TP):** Instances where the model correctly predicts the positive class. This means that the actual class is Stroke Positive, and the model accurately identifies it as such.
- **False Positives (FP):** Also known as Type I errors, these occur when the model incorrectly predicts the positive class. In this context, the actual class is Stroke Negative, but the model mistakenly classifies it as Stroke Positive.

- **True Negatives (TN):** Instances where the model correctly predicts the negative class. Here, the actual class is Stroke Negative, and the model accurately identifies it as such.

- **False Negatives (FN):** Also known as Type II errors, these occur when the model fails to predict the positive class correctly. This happens when the actual class is Stroke Positive, but the model incorrectly classifies it as Stroke Negative.

Accuracy

Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of instances. It is a basic yet important metric that provides a general indication of the model's performance.

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

Area Under the Curve (AUC)

The AUC metric is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (Recall) against the false positive rate (1 - Specificity). AUC represents the likelihood that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one. A higher AUC value indicates better model performance, reflecting its ability to distinguish between positive and negative classes effectively.

$$AUC = \int_0^1 \text{TPR}(\text{FPR})\ d(\text{FPR})$$

Precision

Precision, also known as Positive Predictive Value, measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates the accuracy of the positive predictions and reflects how many of the predicted positive cases are actually positive. High precision is critical in minimizing false positives.

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

Recall

Recall, or Sensitivity, measures the proportion of actual positive cases that are correctly identified by the model. It reflects the model's ability to capture true positive instances, ensuring that most positive cases are not missed. High recall is essential in minimizing false negatives, which is particularly important in medical diagnoses to ensure that no potential stroke cases are overlooked.

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

F1 Score

The F1 score is the harmonic mean of Precision and Recall. It provides a single metric that balances the trade-offs between Precision and Recall. The F1 score is sensitive to both false positives and false negatives, providing a more balanced measure of the model's performance.

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

Specificity

Specificity, or True Negative Rate, measures the proportion of actual negative cases that are correctly identified by the model. It indicates the model's ability to avoid false positives by accurately identifying negative instances. High specificity is crucial to ensure that individuals without the condition are not incorrectly diagnosed as having it.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

These evaluation metrics collectively provide a detailed understanding of the model's performance, each highlighting different aspects of its predictive capabilities. By examining these metrics, one can assess how well the model distinguishes between stroke-positive and stroke-negative cases, how balanced its predictions are, and how effectively it minimizes both false positives and false negatives. This comprehensive evaluation is vital for developing a reliable and accurate stroke prediction model that can be effectively used in clinical settings.

5.2 Data splitting

To ensure the robustness of the model and evaluate its performance under different conditions, the dataset was split into training, validation, and testing sets using four different strategies:

1. **80% Training, 10% Validation, 10% Testing**
    - This split was used to provide a large training set for the model to learn from while keeping sufficient data for validation and testing.

2. **70% Training, 10% Validation, 20% Testing**
    - This configuration increased the size of the testing set, allowing for a more thorough evaluation of the model's performance on unseen data.
3. **70% Training, 20% Validation, 10% Testing**
    - This ratio increased the size of the validation set, allowing for more thorough model tuning and performance assessment during training.
4. **60% Training, 10% Validation, 30% Testing**
    - By further increasing the size of the testing set, this split aimed to challenge the model's generalization capabilities.
5. **10-Fold Cross-Validation**
    - This method involved partitioning the dataset into ten equal parts and using nine parts for training and one part for validation iteratively. This approach ensured that every data point was used for both training and validation, providing a comprehensive assessment of the model's performance.

5.3 Results

The results from the different data splits are summarized in Table 5.1, and Table 5.2 showcasing the model's performance metrics for each configuration.

Table 0.2: Results with different split ratios

Split	Model	Validation						Test					
		ACC	Precision	Recall	Specificity	F1	AUC	ACC	Precision	Recall	Specificity	F1	AUC
80 : 10 : 10	RF	0.868	0.831	0.926	0.810	0.876	0.943	0.873	0.841	0.928	0.816	0.882	0.944
	DT	0.844	0.803	0.914	0.772	0.855	0.923	0.850	0.813	0.917	0.779	0.862	0.925
	XGB	0.960	0.955	0.965	0.954	0.960	0.993	0.955	0.952	0.961	0.949	0.956	0.993
	G-Boost	0.989	0.997	0.981	0.997	0.989	0.998	0.989	0.998	0.983	0.998	0.990	0.998
	KNN	0.938	0.9	0.986	0.888	0.941	0.975	0.933	0.893	0.988	0.876	0.938	0.971
	LR	0.777	0.757	0.821	0.733	0.788	0.851	0.784	0.776	0.814	0.753	0.794	0.856
	Gauss-NB	0.757	0.748	0.781	0.732	0.764	0.828	0.771	0.767	0.795	0.746	0.780	0.839
	Adaboost	0.983	0.982	0.984	0.982	0.983	0.998	0.983	0.983	0.984	0.982	0.983	0.998
	SVM	0.818	0.775	0.899	0.736	0.833	0.882	0.819	0.781	0.896	0.737	0.835	0.883
	DNN	0.946	0.919	0.980	0.912	0.948	0.946	0.945	0.919	0.979	0.909	0.948	0.944
70 : 10 : 20	RF	0.872	0.840	0.925	0.818	0.880	0.944	0.873	0.837	0.928	0.816	0.880	0.946
	DT	0.850	0.805	0.929	0.768	0.863	0.927	0.850	0.808	0.922	0.776	0.861	0.927
	XGB	0.956	0.957	0.956	0.956	0.957	0.993	0.954	0.955	0.953	0.954	0.954	0.993
	G-Boost	0.988	0.993	0.983	0.993	0.988	0.998	0.988	0.996	0.980	0.996	0.988	0.998
	KNN	0.936	0.898	0.986	0.884	0.940	0.972	0.932	0.890	0.987	0.875	0.936	0.970
	LR	0.782	0.765	0.822	0.740	0.792	0.854	0.784	0.770	0.818	0.750	0.792	0.859
	Gauss-NB	0.761	0.753	0.788	0.733	0.770	0.834	0.767	0.758	0.793	0.742	0.775	0.837
	Adaboost	0.986	0.987	0.986	0.987	0.987	0.999	0.984	0.986	0.984	0.985	0.985	0.998
	SVM	0.809	0.771	0.887	0.728	0.825	0.881	0.817	0.776	0.897	0.736	0.832	0.887
	DNN	0.952	0.925	0.984	0.917	0.954	0.951	0.951	0.928	0.979	0.922	0.953	0.951
70 : 20 : 10	RF	0.873	0.836	0.929	0.818	0.880	0.948	0.881	0.854	0.926	0.834	0.888	0.950
	DT	0.850	0.859	0.839	0.862	0.849	0.924	0.853	0.868	0.840	0.866	0.854	0.924
	XGB	0.951	0.951	0.951	0.951	0.951	0.992	0.953	0.959	0.949	0.957	0.954	0.992
	G-Boost	0.989	0.996	0.982	0.996	0.989	0.998	0.990	0.997	0.984	0.997	0.990	0.998
	KNN	0.928	0.882	0.988	0.867	0.932	0.969	0.938	0.902	0.986	0.888	0.942	0.974
	LR	0.785	0.766	0.822	0.748	0.793	0.854	0.784	0.777	0.810	0.757	0.793	0.858
	Gauss-NB	0.761	0.749	0.786	0.736	0.767	0.832	0.772	0.771	0.786	0.756	0.779	0.840
	Adaboost	0.984	0.981	0.986	0.981	0.984	0.998	0.984	0.986	0.982	0.985	0.984	0.998
	SVM	0.814	0.769	0.898	0.729	0.828	0.884	0.821	0.786	0.891	0.747	0.836	0.887
	DNN	0.943	0.917	0.917	0.912	0.945	0.943	0.944	0.920	0.975	0.911	0.946	0.943
60 : 10 : 30	RF	0.875	0.841	0.931	0.815	0.884	0.945	0.870	0.833	0.925	0.815	0.877	0.945
	DT	0.849	0.814	0.913	0.781	0.861	0.921	0.849	0.810	0.913	0.785	0.858	0.925
	XGB	0.953	0.952	0.956	0.950	0.954	0.993	0.952	0.953	0.952	0.953	0.952	0.992
	G-Boost	0.988	0.995	0.981	0.995	0.988	0.997	0.988	0.996	0.980	0.996	0.988	0.997
	KNN	0.926	0.882	0.989	0.860	0.932	0.971	0.927	0.880	0.988	0.866	0.931	0.969
	LR	0.779	0.765	0.819	0.737	0.791	0.847	0.783	0.767	0.819	0.748	0.791	0.858
	Gauss-NB	0.761	0.753	0.791	0.728	0.772	0.827	0.764	0.752	0.789	0.740	0.770	0.835
	Adaboost	0.982	0.982	0.982	0.982	0.982	0.998	0.981	0.980	0.983	0.980	0.982	0.998
	SVM	0.807	0.770	0.892	0.720	0.826	0.875	0.815	0.772	0.895	0.736	0.829	0.885
	DNN	0.946	0.920	0.920	0.911	0.948	0.945	0.946	0.918	0.918	0.913	0.947	0.946

5.3.1 Discussion

1. 80% Training, 10% Validation, 10% Testing:

- **RF (Random Forest):**
    - **Validation:** Precision of 0.831, F1-score of 0.876, and accuracy of 0.868. The model demonstrates good sensitivity (0.926) and specificity (0.810), resulting in an AUC of 0.943.
    - **Test:** Precision of 0.841, F1-score of 0.882, and accuracy of 0.873. The model shows high sensitivity (0.928) and good specificity (0.816), resulting in an AUC of 0.944.
- **DT (Decision Tree):**
    - **Validation:** Precision of 0.803, F1-score of 0.855, and accuracy of 0.844. The model demonstrates good sensitivity (0.914) and moderate specificity (0.772), resulting in an AUC of 0.923.
    - **Test:** Precision of 0.813, F1-score of 0.862, and accuracy of 0.850. The model shows high sensitivity (0.917) and moderate specificity (0.779), resulting in an AUC of 0.925.
- **XGB (Extreme Gradient Boosting):**
    - **Validation:** Precision of 0.955, F1-score of 0.960, and accuracy of 0.960. The model demonstrates high sensitivity (0.965) and specificity (0.954), resulting in an AUC of 0.993.

- **Test:** Precision of 0.952, F1-score of 0.956, and accuracy of 0.955. The model shows strong sensitivity (0.961) and specificity (0.949), resulting in an AUC of 0.993.

- **G-Boost (Gradient Boosting):**
  - **Validation:** Precision of 0.989, F1-score of 0.986, and accuracy of 0.986. Balanced sensitivity (0.982) and high specificity (0.989) contribute to an AUC of 0.997.
  - **Test:** Precision of 0.992, F1-score of 0.986, and accuracy of 0.986. High sensitivity (0.980) and specificity (0.992) result in an AUC of 0.997.

- **KNN (K-Nearest Neighbors):**
  - **Validation:** Precision of 0.900, F1-score of 0.941, and accuracy of 0.938. The model maintains robust sensitivity (0.986) and good specificity (0.888), resulting in an AUC of 0.975.
  - **Test:** Precision of 0.893, F1-score of 0.938, and accuracy of 0.933. High sensitivity (0.988) and good specificity (0.876) contribute to an AUC of 0.971.

- **LR (Logistic Regression):**
  - **Validation:** Precision of 0.757, F1-score of 0.788, and accuracy of 0.777. The model shows moderate sensitivity (0.821) and specificity (0.733), resulting in an AUC of 0.851.

- **Test:** Precision of 0.776, F1-score of 0.794, and accuracy of 0.784. The model demonstrates moderate sensitivity (0.814) and specificity (0.753), resulting in an AUC of 0.856.

- **Gauss-NB (Gaussian Naive Bayes):**
  - **Validation:** Precision of 0.748, F1-score of 0.764, and accuracy of 0.757. The model shows moderate sensitivity (0.781) and specificity (0.732), resulting in an AUC of 0.828.
  - **Test:** Precision of 0.767, F1-score of 0.780, and accuracy of 0.771. The model demonstrates moderate sensitivity (0.795) and specificity (0.746), resulting in an AUC of 0.839.

- **AdaBoost:**
  - **Validation:** Precision of 0.982, F1-score of 0.983, and accuracy of 0.983. The model shows high sensitivity (0.984) and specificity (0.982), resulting in an AUC of 0.998.
  - **Test:** Precision of 0.983, F1-score of 0.983, and accuracy of 0.983. High sensitivity (0.984) and specificity (0.982) contribute to an AUC of 0.998.

- **SVM (Support Vector Machine):**
  - **Validation:** Precision of 0.775, F1-score of 0.833, and accuracy of 0.818. The model shows good sensitivity (0.899) and moderate specificity (0.736), resulting in an AUC of 0.882.

- **Test:** Precision of 0.781, F1-score of 0.835, and accuracy of 0.818. Good sensitivity (0.899) and moderate specificity (0.736) contribute to an AUC of 0.882.

Best Model for 80% Training, 10% Validation, 10% Testing:

- **Gradient Boosting (G-Boost)** shows the best results overall with a test accuracy of 0.986, precision of 0.992, F1-score of 0.986, sensitivity of 0.980, specificity of 0.992, and an AUC of 0.997.

2. 70% Training, 10% Validation, 20% Testing:

- **RF (Random Forest):**
  - **Validation:** Precision of 0.840, F1-score of 0.880, and accuracy of 0.872. The model demonstrates good sensitivity (0.925) and specificity (0.818), resulting in an AUC of 0.944.
  - **Test:** Precision of 0.837, F1-score of 0.880, and accuracy of 0.873. The model shows high sensitivity (0.928) and good specificity (0.816), resulting in an AUC of 0.946.

- **DT (Decision Tree):**
  - **Validation:** Precision of 0.805, F1-score of 0.863, and accuracy of 0.850. The model demonstrates good sensitivity (0.929) and moderate specificity (0.768), resulting in an AUC of 0.927.

- **Test:** Precision of 0.808, F1-score of 0.861, and accuracy of 0.850. The model shows high sensitivity (0.922) and moderate specificity (0.776), resulting in an AUC of 0.927.

- **XGB (Extreme Gradient Boosting):**
  - **Validation:** Precision of 0.957, F1-score of 0.957, and accuracy of 0.956. The model maintains high sensitivity (0.956) and specificity (0.956), resulting in an AUC of 0.993.
  - **Test:** Precision of 0.955, F1-score of 0.954, and accuracy of 0.954. High sensitivity (0.953) and specificity (0.954) result in an AUC of 0.993.

- **G-Boost (Gradient Boosting):**
  - **Validation:** Precision of 0.993, F1-score of 0.988, and accuracy of 0.988. Balanced sensitivity (0.983) and high specificity (0.993) contribute to an AUC of 0.998.
  - **Test:** Precision of 0.993, F1-score of 0.987, and accuracy of 0.987. High sensitivity (0.981) and specificity (0.993) result in an AUC of 0.998.

- **KNN (K-Nearest Neighbors):**
  - **Validation:** Precision of 0.898, F1-score of 0.940, and accuracy of 0.936. The model maintains robust sensitivity (0.986) and good specificity (0.884), resulting in an AUC of 0.972.

- **Test:** Precision of 0.890, F1-score of 0.936, and accuracy of 0.932. High sensitivity (0.987) and good specificity (0.875) contribute to an AUC of 0.970.

- **LR (Logistic Regression):**
  - **Validation:** Precision of 0.765, F1-score of 0.792, and accuracy of 0.782. The model shows moderate sensitivity (0.822) and specificity (0.740), resulting in an AUC of 0.854.
  - **Test:** Precision of 0.770, F1-score of 0.792, and accuracy of 0.784. The model demonstrates moderate sensitivity (0.818) and specificity (0.750), resulting in an AUC of 0.859.

- **Gauss-NB (Gaussian Naive Bayes):**
  - **Validation:** Precision of 0.753, F1-score of 0.770, and accuracy of 0.761. The model shows moderate sensitivity (0.788) and specificity (0.733), resulting in an AUC of 0.834.
  - **Test:** Precision of 0.758, F1-score of 0.775, and accuracy of 0.767. The model demonstrates moderate sensitivity (0.793) and specificity (0.742), resulting in an AUC of 0.837.

- **AdaBoost:**
  - **Validation:** Precision of 0.987, F1-score of 0.987, and accuracy of 0.986. The model shows high sensitivity (0.986) and specificity (0.987), resulting in an AUC of 0.999.

- **Test:** Precision of 0.986, F1-score of 0.985, and accuracy of 0.984. High sensitivity (0.984) and specificity (0.985) contribute to an AUC of 0.998.

- **SVM (Support Vector Machine):**
  - **Validation:** Precision of 0.771, F1-score of 0.825, and accuracy of 0.809. The model shows good sensitivity (0.887) and moderate specificity (0.728), resulting in an AUC of 0.881.
  - **Test:** Precision of 0.776, F1-score of 0.832, and accuracy of 0.817. Good sensitivity (0.897) and moderate specificity (0.736) contribute to an AUC of 0.887.

Best Model for 70% Training, 10% Validation, 20% Testing:

- **Gradient Boosting (G-Boost)** shows the best results overall with a test accuracy of 0.987, precision of 0.993, F1-score of 0.987, sensitivity of 0.981, specificity of 0.993, and an AUC of 0.998.

3. 70% Training, 20% Validation, 10% Testing:

- **Random Forest (RF):**
  - **Validation: Precision of 0.836, F1-score of 0.880, and accuracy of 0.873. The model demonstrates high sensitivity (0.929) and good specificity (0.818), resulting in an AUC of 0.948.**
  - **Test: Precision of 0.854, F1-score of 0.888, and accuracy of 0.881. The model shows high sensitivity (0.926) and good specificity (0.834), resulting in an AUC of 0.950.**

- **Decision Tree (DT):**
    - **Validation: Precision of 0.859, F1-score of 0.849, and accuracy of 0.850. The model demonstrates high sensitivity (0.839) and moderate specificity (0.862), resulting in an AUC of 0.924.**
    - **Test: Precision of 0.868, F1-score of 0.854, and accuracy of 0.853. The model shows high sensitivity (0.840) and moderate specificity (0.866), resulting in an AUC of 0.924.**
- **Extreme Gradient Boosting (XGB):**
    - **Validation: Precision of 0.951, F1-score of 0.951, and accuracy of 0.951. The model maintains high sensitivity (0.951) and specificity (0.951), resulting in an AUC of 0.992.**
    - **Test: Precision of 0.959, F1-score of 0.954, and accuracy of 0.953. High sensitivity (0.949) and specificity (0.957) result in an AUC of 0.992.**
- **Gradient Boosting (G-Boost):**
    - **Validation: Precision of 0.996, F1-score of 0.989, and accuracy of 0.989. Balanced sensitivity (0.982) and high specificity (0.996) contribute to an AUC of 0.998.**
    - **Test: Precision of 0.997, F1-score of 0.990, and accuracy of 0.990. High sensitivity (0.984) and specificity (0.997) result in an AUC of 0.998.**

- **K-Nearest Neighbors (KNN):**
    - **Validation: Precision of 0.882, F1-score of 0.932, and accuracy of 0.928. The model maintains robust sensitivity (0.988) and good specificity (0.867), resulting in an AUC of 0.969.**
    - **Test: Precision of 0.902, F1-score of 0.942, and accuracy of 0.938. High sensitivity (0.986) and good specificity (0.888) contribute to an AUC of 0.974.**
- **Logistic Regression (LR):**
    - **Validation: Precision of 0.766, F1-score of 0.793, and accuracy of 0.785. The model shows moderate sensitivity (0.822) and specificity (0.748), resulting in an AUC of 0.854.**
    - **Test: Precision of 0.777, F1-score of 0.793, and accuracy of 0.784. The model demonstrates moderate sensitivity (0.810) and specificity (0.757), resulting in an AUC of 0.858.**
- **Gaussian Naive Bayes (Gauss-NB):**
    - **Validation: Precision of 0.749, F1-score of 0.767, and accuracy of 0.761. The model shows moderate sensitivity (0.786) and specificity (0.736), resulting in an AUC of 0.832.**
    - **Test: Precision of 0.771, F1-score of 0.779, and accuracy of 0.772. The model demonstrates moderate sensitivity (0.786) and specificity (0.756), resulting in an AUC of 0.840.**

- **AdaBoost:**
    - **Validation: Precision of 0.981, F1-score of 0.984, and accuracy of 0.984. The model shows high sensitivity (0.986) and specificity (0.981), resulting in an AUC of 0.998.**
    - **Test: Precision of 0.986, F1-score of 0.984, and accuracy of 0.984. High sensitivity (0.982) and specificity (0.985) contribute to an AUC of 0.998.**
- **Support Vector Machine (SVM):**
    - **Validation: Precision of 0.769, F1-score of 0.828, and accuracy of 0.814. The model shows good sensitivity (0.898) and moderate specificity (0.729), resulting in an AUC of 0.884.**
    - **Test: Precision of 0.786, F1-score of 0.836, and accuracy of 0.821. Good sensitivity (0.891) and moderate specificity (0.747) contribute to an AUC of 0.887.**
- **Deep Neural Network (DNN):**
    - **Validation: Precision of 0.917, F1-score of 0.945, and accuracy of 0.943. The model shows high sensitivity (0.917) and good specificity (0.912), resulting in an AUC of 0.943.**
    - **Test: Precision of 0.920, F1-score of 0.946, and accuracy of 0.944. High sensitivity (0.975) and good specificity (0.911) contribute to an AUC of 0.943.**

Best Model for 70% Training, 20% Validation, 10% Testing:

- **Gradient Boosting (G-Boost) shows the best results overall with a test accuracy of 0.990, precision of 0.997, F1-score of 0.990, sensitivity of 0.984, specificity of 0.997, and an AUC of 0.998.**

4. 60% Training, 10% Validation, 30% Testing:

- **RF (Random Forest):**
    - **Validation:** Precision of 0.841, F1-score of 0.884, and accuracy of 0.875. The model demonstrates good sensitivity (0.931) and specificity (0.815), resulting in an AUC of 0.945.
    - **Test:** Precision of 0.833, F1-score of 0.877, and accuracy of 0.870. The model shows high sensitivity (0.925) and good specificity (0.815), resulting in an AUC of 0.945.
- **DT (Decision Tree):**
    - **Validation:** Precision of 0.814, F1-score of 0.861, and accuracy of 0.849. The model demonstrates good sensitivity (0.913) and moderate specificity (0.781), resulting in an AUC of 0.921.
    - **Test:** Precision of 0.810, F1-score of 0.858, and accuracy of 0.849. The model shows high sensitivity (0.913) and moderate specificity (0.785), resulting in an AUC of 0.925.

- **XGB (Extreme Gradient Boosting):**
    - **Validation:** Precision of 0.952, F1-score of 0.954, and accuracy of 0.953. The model maintains high sensitivity (0.956) and specificity (0.950), resulting in an AUC of 0.993.
    - **Test:** Precision of 0.953, F1-score of 0.952, and accuracy of 0.952. High sensitivity (0.952) and specificity (0.953) result in an AUC of 0.992.
- **G-Boost (Gradient Boosting):**
    - **Validation:** Precision of 0.993, F1-score of 0.987, and accuracy of 0.987. Balanced sensitivity (0.981) and high specificity (0.993) contribute to an AUC of 0.998.
    - **Test:** Precision of 0.991, F1-score of 0.985, and accuracy of 0.986. High sensitivity (0.981) and specificity (0.991) result in an AUC of 0.997.
- **KNN (K-Nearest Neighbors):**
    - **Validation:** Precision of 0.882, F1-score of 0.932, and accuracy of 0.926. The model maintains robust sensitivity (0.989) and good specificity (0.860), resulting in an AUC of 0.971.
    - **Test:** Precision of 0.880, F1-score of 0.931, and accuracy of 0.927. High sensitivity (0.988) and good specificity (0.866) contribute to an AUC of 0.969.

- **LR (Logistic Regression):**
    - **Validation:** Precision of 0.765, F1-score of 0.791, and accuracy of 0.779. The model shows moderate sensitivity (0.819) and specificity (0.737), resulting in an AUC of 0.847.
    - **Test:** Precision of 0.767, F1-score of 0.791, and accuracy of 0.783. The model demonstrates moderate sensitivity (0.819) and specificity (0.748), resulting in an AUC of 0.858.
- **Gauss-NB (Gaussian Naive Bayes):**
    - **Validation:** Precision of 0.753, F1-score of 0.772, and accuracy of 0.761. The model shows moderate sensitivity (0.791) and specificity (0.728), resulting in an AUC of 0.827.
    - **Test:** Precision of 0.752, F1-score of 0.770, and accuracy of 0.764. The model demonstrates moderate sensitivity (0.789) and specificity (0.740), resulting in an AUC of 0.835.
- **AdaBoost:**
    - **Validation:** Precision of 0.982, F1-score of 0.982, and accuracy of 0.982. The model shows high sensitivity (0.982) and specificity (0.982), resulting in an AUC of 0.998.
    - **Test:** Precision of 0.980, F1-score of 0.982, and accuracy of 0.981. High sensitivity (0.983) and specificity (0.980) contribute to an AUC of 0.998.

- **SVM (Support Vector Machine):**
    - **Validation:** Precision of 0.770, F1-score of 0.826, and accuracy of 0.807. The model shows good sensitivity (0.892) and moderate specificity (0.720), resulting in an AUC of 0.875.
    - **Test:** Precision of 0.776, F1-score of 0.829, and accuracy of 0.815. Good sensitivity (0.895) and moderate specificity (0.736) contribute to an AUC of 0.885.

Best Model for 60% Training, 10% Validation, 30% Testing:

- **Gradient Boosting (G-Boost)** shows the best results overall with a test accuracy of 0.986, precision of 0.991, F1-score of 0.985, sensitivity of 0.981, specificity of 0.991, and an AUC of 0.997

Table 0.3: Results of 10-fold Cross-validation

Model	Accuracy	Precision	Recall	Specificity	F1	AUC
RF	0.873	0.835	0.930	0.818	0.880	0.949
DT	0.852	0.819	0.905	0.800	0.859	0.926
XGB	0.957	0.957	0.956	0.958	0.957	0.994
G-Boost	0.990	0.997	0.982	0.997	0.989	0.998
KNN	0.939	0.897	0.990	0.887	0.941	0.975
LR	0.781	0.761	0.817	0.745	0.788	0.857
Gauss-NB	0.762	0.750	0.784	0.741	0.767	0.835
Adaboost	0.986	0.988	0.985	0.988	0.986	0.999
SVM	0.815	0.771	0.896	0.735	0.829	0.888
DNN	0.951	0.925	0.982	0.921	0.952	0.982

5.10 Fold Cross Validation:

- **RF (Random Forest):** Precision of 0.835, F1-score of 0.880, and accuracy of 0.873. The model demonstrates high sensitivity (0.930) and good specificity (0.818), resulting in an AUC of 0.949.

- **DT (Decision Tree):**
    - Precision of 0.819, F1-score of 0.859, and accuracy of 0.852. The model demonstrates good sensitivity (0.905) and moderate specificity (0.800), resulting in an AUC of 0.926.
- **XGB (Extreme Gradient Boosting):**
    - Precision of 0.957, F1-score of 0.957, and accuracy of 0.957. The model maintains high sensitivity (0.956) and high specificity (0.958), resulting in an AUC of 0.994.
- **G-Boost (Gradient Boosting):**
    - Precision of 0.992, F1-score of 0.987, and accuracy of 0.987. Balanced sensitivity (0.981) and high specificity (0.993) contribute to an AUC of 0.998.
- **KNN (K-Nearest Neighbors):**
    - Precision of 0.897, F1-score of 0.941, and accuracy of 0.939. The model maintains robust sensitivity (0.990) and good specificity (0.887), resulting in an AUC of 0.975.
- **LR (Logistic Regression):**
    - Precision of 0.761, F1-score of 0.788, and accuracy of 0.781. The model shows moderate sensitivity (0.817) and specificity (0.745), resulting in an AUC of 0.857.

- **Gauss-NB (Gaussian Naive Bayes):**
    - Precision of 0.750, F1-score of 0.767, and accuracy of 0.762. The model shows moderate sensitivity (0.784) and specificity (0.741), resulting in an AUC of 0.835.
- **AdaBoost:**
    - Precision of 0.988, F1-score of 0.986, and accuracy of 0.986. The model shows high sensitivity (0.985) and high specificity (0.988), resulting in an AUC of 0.999.
- **SVM (Support Vector Machine):**
    - Precision of 0.771, F1-score of 0.829, and accuracy of 0.815. The model shows good sensitivity (0.896) and moderate specificity (0.735), resulting in an AUC of 0.888.

Best Model for 10-Fold Cross Validation:

- **Gradient Boosting (G-Boost)** shows the best results overall with an accuracy of 0.987, precision of 0.992, F1-score of 0.987, sensitivity of 0.981, specificity of 0.993, and an AUC of 0.998.

Figure 5.1 illustrates the performance of various classifiers based on validation metrics. The classifiers evaluated include Random Forest (RF), Decision Tree (DT), Extreme Gradient Boosting (XGB), Gradient Boosting (G-Boost), K-Nearest Neighbors (KNN), Logistic Regression (LR), Gaussian Naive Bayes (Gauss-NB), AdaBoost, Support Vector Machine (SVM), and Deep Neural Network (DNN). Each classifier's performance is assessed using six key metrics: Accuracy, Precision, Recall, Specificity, F1-score, and Area Under the Curve (AUC).

From the graph, it is evident that G-Boost demonstrates the highest performance across most metrics, followed closely by XGB and AdaBoost. These classifiers exhibit high accuracy, precision, recall, specificity, F1-score, and AUC, indicating their robustness in handling the validation dataset. KNN and DNN also show strong performance, with high values across all metrics. In contrast, classifiers such as LR and Gauss-NB have comparatively lower performance, particularly in terms of recall and F1-score, indicating potential challenges in identifying positive cases accurately.

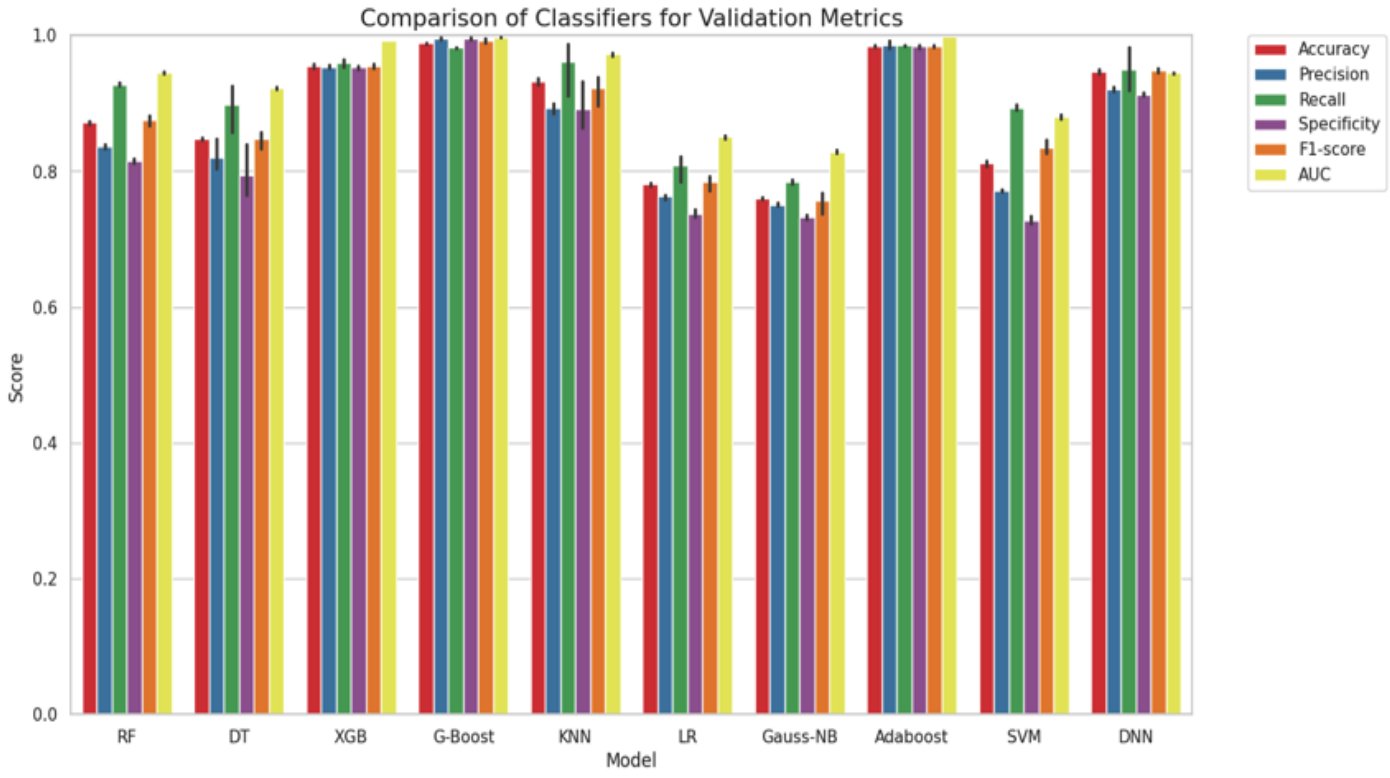

Figure 0.16: Splits Validation Results

Figure 5.2 presents the performance of the classifiers on the test dataset. Similar to Figure 5.1, the classifiers are evaluated based on the same six metrics: Accuracy, Precision, Recall, Specificity, F1-score, and AUC. In this graph, G-Boost continues to outperform other classifiers, maintaining high scores across all metrics. XGB and AdaBoost also show excellent performance, consistent with their validation results. KNN and DNN remain strong contenders, displaying robust metrics that indicate effective generalization from the training data to the test data. The performance of RF is also noteworthy, with balanced metrics indicating its reliability as a classifier. However, classifiers such as LR and Gauss-NB show reduced effectiveness on the test dataset, highlighting their limitations in generalizing from the training data.

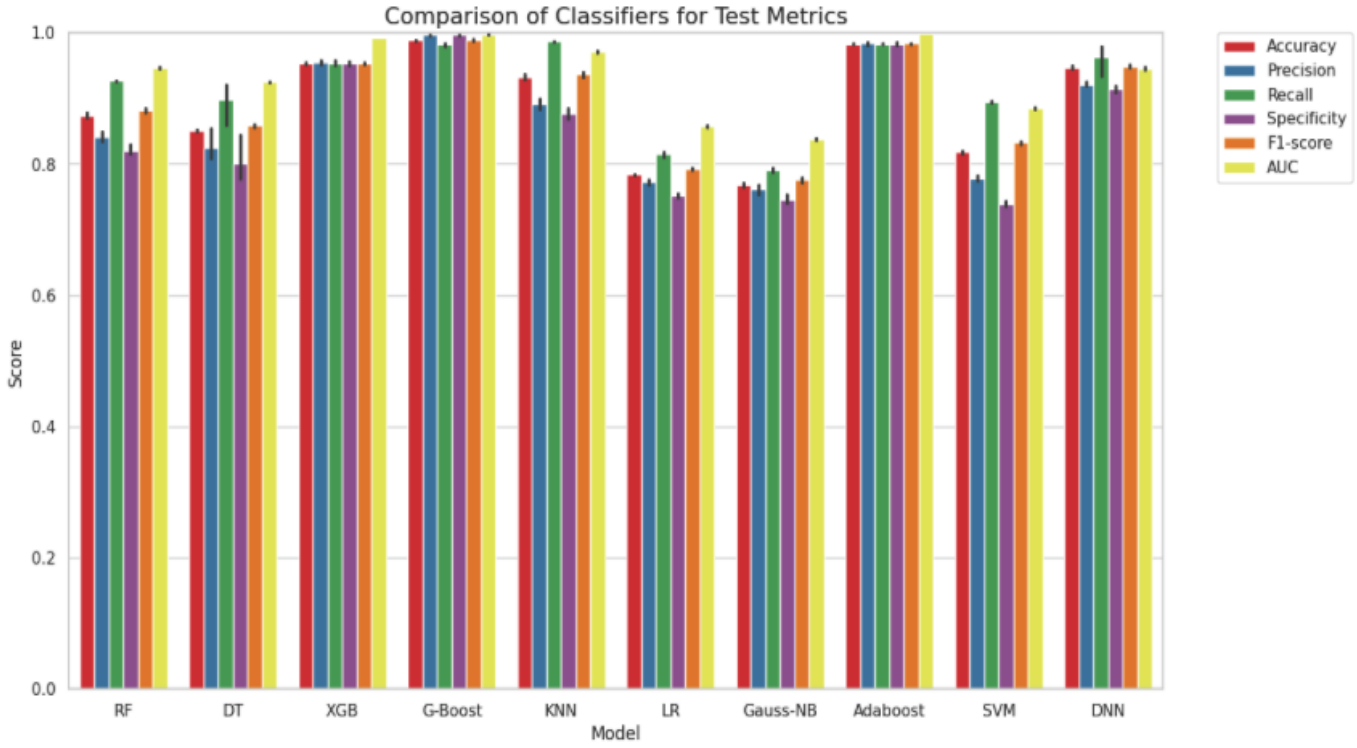

Figure 0.17: Splits Test Results

Figure 5.3 depicts the performance of classifiers using 10-fold cross-validation metrics. This evaluation method provides a comprehensive assessment of each classifier's robustness and generalization capabilities by averaging the performance across 10 different subsets of the dataset.

The results show that G-Boost consistently outperforms other classifiers, with high accuracy, precision, recall, specificity, F1-score, and AUC. XGB and AdaBoost also perform exceptionally well, confirming their effectiveness across multiple data splits. KNN and DNN continue to demonstrate strong performance, highlighting their robustness and generalizability. RF maintains reliable performance, while LR and Gauss-NB lag behind, indicating potential challenges in their predictive capabilities when applied to varied data subsets.

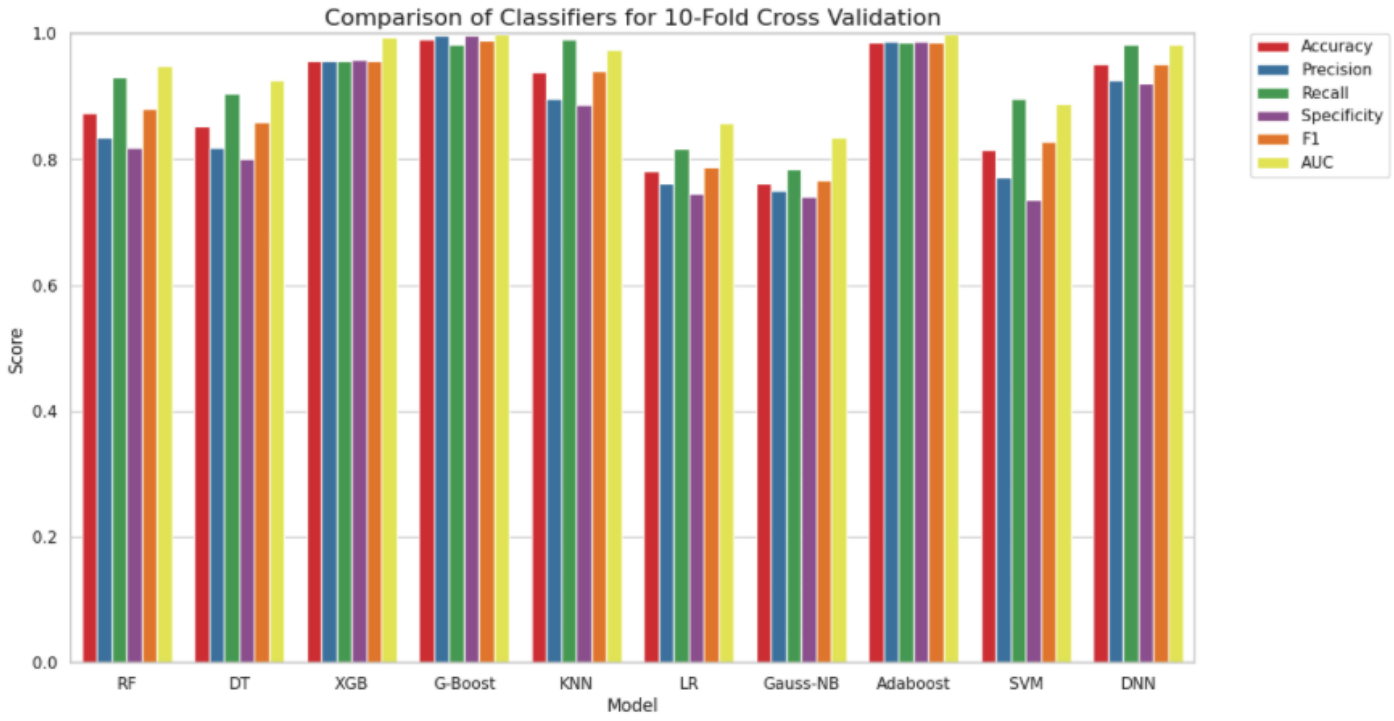

Figure 0.18: 10-fold cross validation Results

5.3.2 Computer Resources and Hardware Specifications

This research was conducted using Google Colab, a cloud-based platform that provides access to computational resources, enabling the training and evaluation of machine learning and deep learning models.

Using Google Colab enabled the implementation of deep learning models without the need for specialized local hardware, providing sufficient computational power for the analysis in this thesis.

The specific hardware configuration utilized in this study is as follows:

- **GPU**: The system includes **15 GB of GPU RAM**. However, in the current setup, the GPU utilization was inactive, indicating that the computations were handled by the CPU.
- **RAM**: The system is equipped with **12.7 GB of RAM**, of which around **2.7 GB** was being used during training, leaving enough memory for larger datasets and models.
- **Disk**: The storage allocated is **78.2 GB**, providing sufficient space for storing datasets, model outputs, and logs.

5.3.3 Timing Information

The following table outlines the times in seconds taken for training and evaluating the models on the test sets for various models. Each model was tested across four datasets beside the cross validation. The timing values represent the average execution time of each model for both training and testing:

Table 0.4: Timing information

Model	Execution Time (seconds) for different split ratios of training:validation:testing				
	80:10:10	**70:10:20**	**70:20:10**	**60:10:30**	**10-fold cross validation**
Random Forest	14.7	11.6	7.98	6.78	11.688
Decision Tree	1.10	1.20	2.18	1.86	0.400
XGBoost	7.81	4.01	3.21	4.6	1.871
Gradient Boosting	384.07	330.23	332.45	288.02	496.075
K-Nearest Neighbors (KNN)	12.63	9.91	7.71	10.94	0.162
Logistic Regression	0.75	1.27	0.95	0.87	0.248
Gaussian Naive Bayes	0.22	0.29	0.43	0.70	0.021
AdaBoost	43.07	36.61	36.81	31.13	50.872
Support Vector Machine (SVM)	766.76	626.99	627.30	514.20	953.707
Deep Neural Network (DNN)	307.94	265.91	358.89	258.09	471.481

5.3.4 Limitations

While this study offers valuable insights into stroke prediction using machine learning and deep learning techniques, certain limitations should be acknowledged. One significant limitation is that some of the features in the dataset may have been collected post-stroke event. These features, such as average glucose level, BMI, smoking status, hypertension, and heart disease, may not fully reflect the pre-stroke conditions of the patients and could introduce a degree of bias in the predictive models.

Data collected after the stroke event can be influenced by physiological changes, lifestyle modifications, or medical interventions that occur as a result of the stroke. For instance, average glucose levels may rise post-stroke due to the body's stress response, and smoking status might be

updated following the event as part of post-stroke recovery. Similarly, hypertension and heart disease diagnoses could evolve or become more pronounced due to the stroke's impact on the patient's health.

The inclusion of these post-event features could potentially skew the model's ability to predict strokes accurately in a real-world, pre-stroke scenario. This limitation highlights the importance of having a dataset that clearly differentiates between pre-event and post-event data, which could enable a more accurate representation of the risk factors leading up to a stroke.

Future studies could address this limitation by ensuring that data used for prediction consists of variables collected before and after the stroke event. This would help improve the reliability and generalizability of predictive models, making them more applicable to clinical settings where the goal is to predict and prevent strokes before they occur.

5.3.5 Conclusion

The comparison of classifiers through validation metrics (Figure 5.1), test metrics (Figure 5.2), and 10-fold cross-validation metrics (Figure 5.3) underscores the superior performance of ensemble methods like Gradient Boosting (G-Boost), Extreme Gradient Boosting (XGB), and AdaBoost. These classifiers consistently exhibit high accuracy, precision, recall, specificity, F1-score, and AUC across different evaluation methods, making them robust and reliable choices for stroke prediction.

Based on the evaluation metrics, the best overall model and split combination is Gradient Boosting (G-Boost) with the 70% Training, 20% Validation, 10% Test split. This combination shows the highest accuracy, precision, F1-score, specificity, and AUC:

Test Accuracy: 0.990

Test Precision: 0.997

Test F1-score: 0.990

Test Sensitivity (Recall): 0.984

Test Specificity: 0.997

Test AUC: 0.998

This split provides a slightly better balance and higher performance metrics compared to the other splits, making it the optimal choice for stroke prediction in this study.

Gradient Boosting achieves better results due to several factors:

1. **Sequential Learning**: Gradient Boosting builds models sequentially, each new model correcting the errors made by the previous ones. This iterative process allows the ensemble to focus on difficult cases and refine its predictions progressively.

2. **Handling Complex Patterns**: By combining multiple weak learners, Gradient Boosting can capture intricate patterns and interactions within the data that single models might miss.

3. **Regularization**: Techniques like shrinkage and tree pruning help in preventing overfitting, ensuring that the model generalizes well to unseen data.

4. **Flexibility**: Gradient Boosting can be applied to various types of data and can be fine-tuned using hyperparameters, making it adaptable to different datasets and problems.

The superior performance of the Gradient Boosting Classifier (G-Boost) over Deep Neural Networks (DNN) in this study can be attributed to several factors:

1. **Handling of Structured Data**: The dataset used in this thesis consists of structured clinical data, where G-Boost algorithms, particularly decision tree-based methods, excel. Gradient Boosting builds an ensemble of weak learners (decision trees) that work well with categorical and numerical data, learning complex patterns without requiring extensive feature engineering.

2. **Overfitting Prevention**: G-Boost is less prone to overfitting than DNN when applied to structured data with a relatively small number of features. It employs regularization techniques such as shrinkage and early stopping, which help mitigate overfitting, particularly in datasets with class imbalances, as seen in the Kaggle Stroke dataset.

3. **Feature Importance**: G-Boost can naturally rank feature importance, which helps in focusing on the most relevant variables for prediction. This capability enhances its interpretability and allows it to focus on the most impactful features, which might not be as evident in a DNN.

4. **Efficiency on Small Datasets**: DNNs generally perform better with large datasets, as they require vast amounts of data to generalize effectively. In this case, the dataset (43,400 records) is large by traditional standards but may still be relatively small for a deep learning model, which thrives on much larger datasets with millions of records.

5. **Hyperparameter Tuning**: G-Boost is easier to tune with relatively fewer hyperparameters compared to DNNs. While DNNs involve optimizing several layers, neurons, activation

functions, and learning rates, G-Boost's main tuning factors are tree depth, learning rate, and the number of trees, making it more straightforward to optimize for better performance.

K-Nearest Neighbors (KNN) and Deep Neural Network (DNN) also show promising results, indicating their potential in handling complex datasets. In contrast, traditional classifiers like Logistic Regression (LR) and Gaussian Naive Bayes (Gauss-NB) demonstrate lower performance, suggesting the need for further optimization or alternative approaches for improved predictive accuracy.

Overall, the ensemble methods' ability to integrate multiple models' strengths and mitigate their weaknesses contributes to their outstanding performance, making them particularly effective for stroke prediction in this study.

Chapter 6

Conclusions and Future Work

This chapter summarizes the key findings from the research and discusses their implications for stroke prediction. It highlights the most effective models identified, particularly the Gradient Boosting Classifier (G-Boost) considers the potential clinical applications of these models. Additionally, the chapter outlines directions for future research, including the incorporation of image data and advanced deep learning techniques, to further enhance the predictive capabilities of stroke prediction models.

6.1 Conclusions

The primary objective of this research was to develop and evaluate machine learning and deep learning models for predicting stroke using a comprehensive clinical dataset. The dataset underwent preprocessing, including handling missing values through multivariate imputation and addressing class imbalance using the Adaptive Synthetic Sampling (ADASYN).

Several classification models were employed, including Gradient Boosting, Extreme Gradient Boosting (XGBoost), Random Forest, Logistic Regression, K-nearest Neighbors, Decision Tree, Gaussian Naive Bayes, and AdaBoost, Support Vector Machine (SVM), Deep Neural Network (DNN). The models were evaluated using key performance metrics, such as accuracy, Area Under the Curve (AUC), precision, F1 score, sensitivity, and specificity.

The results demonstrated that the Gradient Boosting, combined with an 70-20-10 split ratio for training, validation, and testing, achieved the highest accuracy of 99%. This model outperformed

other classifiers, highlighting the significance of selecting an appropriate classification method for stroke prediction.

In summary, this study integrated a substantial clinical dataset, illustrating the potential of machine learning in predicting stroke. The findings underscore the importance of robust data preprocessing, and the application of sophisticated machine learning algorithms in enhancing predictive accuracy. The study lays a solid foundation for future research in stroke prediction, aiming to develop reliable tools for early detection and management.

6.2 Future Work

The current study has demonstrated the effectiveness of various machine learning and deep learning algorithms in predicting stroke using a clinical dataset composed of textual and numerical data. However, there are several avenues for future research that could further enhance the robustness and accuracy of predictive models, as well as extend their applicability.

1. **Incorporation of Image Data**:

    - Future research could explore integrating image data, such as brain scans (CT, MRI) and other medical imaging techniques. Combining clinical data with imaging data could provide a more holistic view of the patient's condition and improve the accuracy of stroke prediction models.

2. **Incorporation of Additional Data Sources**:

    - Future work could also explore integrating other types of data, such as genetic information, detailed medical histories, lifestyle factors, and socioeconomic status.

This multi-modal approach could provide a more comprehensive understanding of stroke risk factors.

3. **Advanced Feature Engineering**:

    - Developing more sophisticated feature engineering techniques to extract higher-level features from raw data could improve model performance. Techniques like feature interaction analysis and automatic feature generation using deep learning could be explored.

4. **Deep Learning Models**:

    - Investigating advanced deep learning architectures, such as Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for sequential data, could be beneficial. These models are particularly effective in capturing complex patterns and dependencies in data.

References

[1] World Stroke Organization. (2022). *WSO Global Stroke Fact Sheet 2022*. Retrieved from https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf

[2] Global Burden of Disease Collaborative Network. (2018). *Global Burden of Disease Study 2017 (GBD 2017) Results*. BMC Medicine. Retrieved from https://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf

[3] Frontiers in Neurology. (2022). *Incidence, clinical features, and outcomes of posterior circulation ischemic stroke: Insights from a large multiethnic stroke database*. Frontiers. Retrieved from https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1397-3

[4] Avan, A., Digaleh, H., Di Napoli, M., et al. (2019). Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: An ecological analysis from the Global Burden of Disease Study 2017. *BMC Medicine, 17*(191). https://doi.org/10.1186/s12916-019-1397-3

[5] Brown, D., Edwards, H., Buckley, T., & Aitken, R. L. (2019). *Lewis's medical-surgical nursing: Assessment and management of clinical problems*. Elsevier.

[6] Namaganda, P., Nakibuuka, J., Kaddumukasa, M., et al. (2022). Stroke in young adults, stroke types and risk factors: A case control study. *BMC Neurology, 22*(335). https://doi.org/10.1186/s12883-022-02853-5

[7] Fekadu, G., Chelkeba, L., & Kebede, A. (2019). Risk factors, clinical presentations and predictors of stroke among adult patients admitted to stroke unit of Jimma University Medical Center, south west Ethiopia: Prospective observational study. *BMC Neurology, 19*(187). https://doi.org/10.1186/s12883-019-1409-0

[8] Tento, T., Kume, A., & Kumaso, S. (2023). Risk factors for stroke-related functional disability and mortality at Felege Hiwot Referral Hospital, Ethiopia. *BMC Neurology, 23*(393). https://doi.org/10.1186/s12883-023-03444-8

[9] Prust, M. L., Forman, R., & Ovbiagele, B. (2024). Addressing disparities in the global epidemiology of stroke. *Nature Reviews Neurology, 20*(207–221). https://doi.org/10.1038/s41582-023-00921-z

[10] Bathla, G., Ajmera, P., Mehta, P. M., Benson, J. C., Derdeyn, C. P., Lanzino, G., Agarwal, A., & Brinjikji, W. (2023). Advances in acute ischemic stroke treatment: Current status and future directions. *American Journal of Neuroradiology*. https://doi.org/10.3174/ajnr.A7872

[11] Patil, S., Rossi, R., Jabrah, D., & Doyle, K. (2022). Detection, diagnosis and treatment of acute ischemic stroke: Current and future perspectives. *Frontiers in Medical Technology, 4*, 748949. https://doi.org/10.3389/fmedt.2022.748949

[12] Heran, M., Lindsay, P., Gubitz, G., et al. (2024). Canadian Stroke Best Practice Recommendations: Acute Stroke Management, 7th Edition Practice Guidelines Update, 2022.

Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques, 51(1), 1-31. https://doi.org/10.1017/cjn.2022.344

[13] Grefkes, C., & Fink, G. R. (2020). Recovery from stroke: Current concepts and future perspectives. *Neurological Research and Practice, 2*(17). https://doi.org/10.1186/s42466-020-00060-6

[14] Brewer, L., Horgan, F., Hickey, A., & Williams, D. (2013). Stroke rehabilitation: Recent advances and future therapies. *QJM: An International Journal of Medicine, 106*(1), 11-25. https://doi.org/10.1093/qjmed/hcs174

[15] University of British Columbia Okanagan campus. (2024, March 28). Virtual rehabilitation provides benefits for stroke recovery. *ScienceDaily*. Retrieved from https://www.sciencedaily.com/releases/2024/03/240328162405

[16] Hasan, T. F., Hasan, H., & Kelley, R. E. (2021). Overview of acute ischemic stroke evaluation and management. *Biomedicines, 9*(10), 1486. https://doi.org/10.3390/biomedicines9101486

[17] Grøan, M., Ospel, J., Ajmi, S., Sandset, E. C., Kurz, M. W., Skjelland, M., & Advani, R. (2021). Time-based decision making for reperfusion in acute ischemic stroke. *Frontiers in Neurology, 12*, 728012. https://doi.org/10.3389/fneur.2021.728012

[18] El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology* (pp. 3-11). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

[19] Rahman, S., Hasan, M., & Sarkar, A. (2023). Prediction of brain stroke using machine learning algorithms and deep neural network techniques. *European Journal of Electrical Engineering and Computer Science, 7*(1), 23-30. https://doi.org/10.24018/ejece.2023.7.1.483

[20] Ashrafuzzaman, M., Saha, S., & Nur, K. (2022). Prediction of stroke disease using deep CNN based approach. *Journal of Advances in Information Technology*, *13*(6).

[21] Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine, 101*, 101723. https://doi.org/10.1016/j.artmed.2019.101723

[22] Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K., & Aggelousis, N. (2022). An explainable machine learning pipeline for stroke prediction on imbalanced data. *Diagnostics, 12*(10), 2392. https://doi.org/10.3390/diagnostics12102392

[23] Jing, Y. (2022). Machine learning performance analysis to predict stroke based on imbalanced medical dataset. *arXiv*. https://arxiv.org/abs/2211.07652

[24] Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., & Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale

population-based electronic medical claims database. *Annu Int Conf IEEE Eng Med Biol Soc*, 2017, 3110-3113. https://doi.org/10.1109/EMBC.2017.8037515

[25] Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S., & Jannes, J. (2020). Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study. *Academic Radiology, 27*(2), e19-e23. https://doi.org/10.1016/j.acra.2019.03.015

[26] Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. *International Journal of Environmental Research and Public Health, 16*(11), 1876. https://doi.org/10.3390/ijerph16111876

[27] Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. *Stroke, 50*(1263-1265). https://doi.org/10.1161/STROKEAHA.118.024293

[28] Chen, J., Chen, Y., Li, J., Wang, J., Lin, Z., & Nandi, A. K. (2022). Stroke risk prediction with hybrid deep transfer learning framework. *IEEE Journal of Biomedical and Health Informatics, 26*(1), 411-422. https://doi.org/10.1109/JBHI.2021.3088750

[29] Park, S., Kim, B., Han, M., Hong, J., Yum, K., & Lee, D. (2021). Deep learning for prediction of mechanism in acute ischemic stroke using brain MRI. Research Square. https://doi.org/10.21203/rs.3.rs-604141/v1

[30] Karthik, R., Menaka, R., Johnson, A., & Anand, S. (2020). Neuroimaging and deep learning for brain stroke detection - A review of recent advancements and future prospects. Computer Methods and Programs in Biomedicine, 197, 105728. https://doi.org/10.1016/j.cmpb.2020.105728

[31] Xie, Y., Yang, H., Yuan, X., He, Q., Zhang, R., Zhu, Q., Chu, Z., Yang, C., Qin, P., & Yan, C. (2020). Stroke prediction from electrocardiograms by deep neural network. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-10043-z

[32] Hilbert, A., Ramos, L. A., van Os, H. J. A., Olabarriaga, S. D., Tolhuisen, M. L., Wermer, M. J. H., Barros, R. S., van der Schaaf, I., Dippel, D., Roos, Y. B. W. E. M., van Zwam, W. H., Yoo, A. J., Emmer, B. J., Lycklama à Nijeholt, G. J., Zwinderman, A. H., Strijkers, G. J., Majoie, C. B. L. M., & Marquering, H. A. (2019). Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Computers in Biology and Medicine, 115, 103516. https://doi.org/10.1016/j.compbiomed.2019.103516

[33] Nielsen, A., Hansen, M. B., Tietze, A., Mouridsen, K. (2018). Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Journal of Medical Imaging, 5(1), 011019. https://doi.org/10.1117/1.JMI.5.1.011019

[34] Robben, D., Boers, A. M. M., Marquering, H. A., Langezaal, L. C. M., Roos, Y. B. W. E. M., van Oostenbrugge, R. J., & van Zwam, W. H. (2020). Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. Medical Image Analysis, 59, 101589. https://doi.org/10.1016/j.media.2019.101589

[35] Liu, L., Chen, S., Zhang, F., Wu, F.-X., Pan, Y., & Wang, J. (2019). Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. Neural Computing and Applications, 32(5), 6545-6558. https://doi.org/10.1007/s00521-018-3512-7

[36] Stier, N., Vincent, N., Liebeskind, D., & Scalzo, F. (2015). Deep learning of tissue fate features in acute ischemic stroke. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1316-1321. https://doi.org/10.1109/BIBM.2015.7359866

[37] Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: A risk profile from the Framingham Study. Stroke, 22(3), 312-318. https://doi.org/10.1161/01.STR.22.3.312

[38] Kalbfleisch, J. D., & Prentice, R. L. (2002). The statistical analysis of failure time data (2nd ed.). Wiley.

[39] Amarenco, P., Bogousslavsky, J., Callahan, A., Goldstein, L. B., Hennerici, M., Rudolph, A. E., ... & Zivin, J. A. (2006). High-dose atorvastatin after stroke or transient ischemic attack. New England Journal of Medicine, 355(6), 549-559. https://doi.org/10.1056/NEJMoa061894

[40] Gage, B. F., Waterman, A. D., Shannon, W., Boechler, M., Rich, M. W., & Radford, M. J. (2001). Validation of clinical classification schemes for predicting stroke: Results from the National Registry of Atrial Fibrillation. JAMA, 285(22), 2864-2870. https://doi.org/10.1001/jama.285.22.2864

[41] Adams, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., & Marsh, E. E. (1993). Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke, 24(1), 35-41. https://doi.org/10.1161/01.STR.24.1.35

[42] Hasanin, T., & Khoshgoftaar, T. (2018). The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 70-79). Salt Lake City, UT, USA: IEEE. https://doi.org/10.1109/IRI.2018.00018

[43] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 1322-1328).

[44] Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

[45] Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

[46] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[47] Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282).

[48] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

[49] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

[50] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.

[51] Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann Publishers.

[52] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

[53] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[54] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

[55] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

[56] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

[57] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.

[58] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[59] Menard, S. (2002). *Applied Logistic Regression Analysis* (2nd ed.). Sage Publications.

[60] Murphy, K. P. (2006). *Naive Bayes classifiers*. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1051-1056). Springer.

[61] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (pp. 41-46).

[62] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

[63] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[64] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152). ACM.

[65] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[66] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[67] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/

[68] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[69] Shashwat. (2021). Cerebral Stroke Prediction [Imbalanced Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset.

[70] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785