OPTIMIZING ISCHEMIC STROKE CLASSIFICATION USING MACHINE LEARNING FOR CLINICAL APPLICABILITY AT BANJARMASIN HOSPITALS

AGUS BYNA

SULTAN IDRIS EDUCATION UNIVERSITY
2025

OPTIMIZING ISCHEMIC STROKE CLASSIFICATION USING MACHINE LEARNING FOR CLINICAL APPLICABILITY AT BANJARMASIN HOSPITALS

AGUS BYNA

THESIS PRESENTED TO QUALIFY FOR A DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTING AND META-TECHNOLOGY
SULTAN IDRIS EDUCATION UNIVERSITY
2025

UPSI/IPS-3/BO 32
Pind : 00 m/s: 1/1

Please tick (√)	
Project Paper	
Masters by Research	
Master by Mixed Mode	
PhD	√

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the 5 (day) day of August (month) 2025 (year).

i. Student's Declaration:

I, AGUS BYNA (MATRIC NO. P20211002729), Faculty of Computing and Meta-Technology hereby declare dissertation / thesis for Doctor of Philosophy in Artificial Intelligence titled OPTIMIZING ISCHEMIC STROKE CLASSIFICATION USING MACHINE LEARNING FOR CLINICAL APPLICABILITY AT BANJARMASIN HOSPITAL is my original work. I have not plagiarised from any other scholar's work and any sources that contains copyright had been cited properly for the permitted meanings. Any quotations, excerpt, reference or re-publication from or any works that has copyright had been clearly and well cited.

Signature of the student

ii. Supervisor Declaration:

I, Prof. Madya Ts. Dr. Muhammad Modi bin Lakulu hereby certify that the work entitled OPTIMIZING ISCHEMIC STROKE CLASSIFICATION USING MACHINE LEARNING FOR CLINICAL APPLICABILITY AT BANJARMASIN HOSPITAL was prepared by the above the named student, and was submitted to the Institute of Graduate Studies as a partial fulfilment for the conferment of Doctor of Philosophy in Artificial Intelligence, and the aforementioned work, to the best of my knowledge, is said student's work.

5/08/2025
Date

Prof. Madya Ts Dr. Muhammad Modi Lakulu
Pensyarah
Fakulti Komputeran dan Meta-Teknologi
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak

Signature of the Supervisor

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: OPTIMIZING ISCHEMIC STROKE CLASSIFICATION USING MACHINE LEARNING FOR CLINICAL APPLICABILITY AT BANJARMASIN HOSPITALS

No. Matrik /*Matric's No.*: P20211002729

Saya / *I* : AGUS BYNA

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah */ Please tick (√) for category below:-*

☐	**SULIT/*CONFIDENTIAL***	Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*
☐	**TERHAD/*RESTRICTED***	Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restircted information as specified by the organization where research was done.*
☑	**TIDAK TERHAD / *OPEN ACCESS***	

(Tandatangan Pelajar/ Signature)

Tarikh: 05/08/2025

Prof. Madya Ts Dr. Muhammad Modi Lakulu
Pensyarah
Fakulti Komputeran dan Meta-Teknologi
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak

(Tandatangan Penyelia / *Signature of Supervisor*)
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Catatan: Jika Tesis/Disertasi ini **SULIT** @ **TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGEMENT

All praise is due to Allah SWT, the Lord of the universe and the owner of all knowledge, who has granted the author the strength and wisdom to complete this thesis entitled "Optimizing Ischemic Stroke Classification Using Machine Learning for Clinical Applicability at Banjarmasin Hospitals." The journey has been marked by invaluable support and guidance from many parties, to whom the author extends heartfelt gratitude. The author wishes to express sincere appreciation to: Assoc. Prof. TS. Dr. Muhammad Modi Bin Lakulu, the first supervisor, for his unwavering patience and guidance throughout the three-year research period. His insights and dedication have been instrumental in shaping this study. The Dean of the Faculty, for facilitating administrative processes, providing exceptional educational resources at UPSI, and extending remarkable hospitality to support academic excellence. The Vice Dean, for continuous motivation and for fostering collaboration within the field of Computer Education. The esteemed lecturers, for their profound knowledge and clear instruction in data structures, algorithms, and other foundational subjects crucial to this research. All research experts, who generously contributed data and shared specialized knowledge in support of this work. Faculty members and staff, for their thoughtful assistance, encouragement, and prayers, which have greatly supported the completion of this thesis. The author also gratefully acknowledges the contributions of individuals who are not mentioned by name but whose support has been deeply appreciated. May Allah SWT reward them manifold for their kindness. This research is acknowledged to have limitations, and the author warmly welcomes constructive criticism, suggestions, and insights to improve future studies. It is sincerely hoped that this thesis contributes meaningfully to the ongoing development of scientific knowledge and clinical practice.

ABSTRACT

The increasing prevalence of ischemic stroke—particularly in Banjarmasin, Indonesia—demands the development of accurate, robust, and interpretable classification models to support timely and effective clinical decision-making. Conventional approaches and standard machine learning techniques often fall short when addressing the challenges posed by highly imbalanced medical datasets (91.40% majority vs. 8.60% minority) and limited model transparency, both of which impede clinical adoption. To overcome these limitations, this study introduces a rigorously optimized framework based on the XGBoost algorithm, enhanced by the Synthetic Minority Over-sampling Technique (SMOTE) to correct for class imbalance. The methodology incorporates a structured Train-Validation-Test split, 10-fold cross-validation, and performance assessment using mean (μ) and standard deviation (σ). Two hyperparameter tuning strategies were implemented, with Random Forest employed as a comparative benchmark. SHapley Additive exPlanations (SHAP) were integrated to improve model interpretability. The XGBoost Hyperparameter Tuning Type 1 model, supported by Enhanced SMOTE, achieved a mean classification accuracy of 99.007% (±0.14%) and consistently exhibited high sensitivity (>97%) in detecting the minority class. Both ensemble models—XGBoost and Random Forest—significantly outperformed the Decision Tree classifier, with no notable performance discrepancy between them. SHAP analysis consistently identified hypertension, heart disease, and genetic predisposition as key features contributing to classification outcomes. This research presents a robust and transparent machine learning framework for ischemic stroke classification, offering clinically relevant insights to aid in risk stratification and targeted intervention. The integration of SHAP enhances model explainability, thereby promoting greater trust among clinicians and informing improved strategies for stroke prevention and management in Banjarmasin.

MENGOPTIMALKAN KLASIFIKASI STROKE ISKEMIK MENGGUNAKAN PEMBELAJARAN MESIN UNTUK KEBOLEHAN KLINIK DI HOSPITAL BANJARMASIN

ABSTRAK

Peningkatan prevalensi strok iskemik—terutamanya di Banjarmasin, Indonesia—menuntut pembangunan model klasifikasi yang tepat, kukuh, dan boleh ditafsirkan untuk menyokong pembuatan keputusan klinikal yang tepat pada masanya dan berkesan. Pendekatan konvensional dan teknik pembelajaran mesin standard sering kali tidak mencukupi dalam menangani cabaran yang ditimbulkan oleh set data perubatan yang sangat tidak seimbang (91.40% majoriti vs. 8.60% minoriti) dan ketelusan model yang terhad, kedua-duanya menghalang penerimaan klinikal. Untuk mengatasi kekangan ini, kajian ini memperkenalkan rangka kerja yang dioptimumkan dengan teliti berdasarkan algoritma XGBoost, yang dipertingkatkan oleh Teknik Sintetik Minoriti *Over-sampling* (*SMOTE*) untuk membetulkan ketidakseimbangan kelas. Metodologi ini menggabungkan pembahagian terstruktur *Train-Validation-Test*, 10-lipat silang validasi, dan penilaian prestasi menggunakan min (μ) dan sisihan piawai (σ). Dua strategi penyetelan hiperparameter telah dilaksanakan, dengan *Random Forest* digunakan sebagai penanda aras perbandingan. *SHapley Additive exPlanations* (*SHAP*) telah disepadukan untuk meningkatkan kebolehcapaian model. Model Penalaan Hiperparameter XGBoost Jenis 1, yang disokong oleh pertingkatkan SMOTE, mencapai ketepatan klasifikasi purata sebanyak 99.007% (±0.14%) dan secara konsisten menunjukkan kepekaan tinggi (>97%) dalam mengesan kelas minoriti.. Kedua-dua model ensembel *XGBoost* dan *Random Forest* secara signifikan mengatasi pengelasan *Decision Tree*, tanpa sebarang perbezaan prestasi yang ketara antara mereka. Analisis *SHAP* secara konsisten mengenal pasti hipertensi, penyakit jantung, dan predisposisi genetik sebagai ciri-ciri utama yang menyumbang kepada hasil klasifikasi. Penyelidikan ini mempersembahkan rangka kerja pembelajaran mesin yang kukuh dan telus untuk pengelasan strok iskemia, menawarkan pandangan yang relevan secara klinikal untuk membantu dalam pengkelasan risiko dan intervensi yang disasarkan. Pengintegrasian *SHAP* meningkatkan kebolehjelaskan model, sekali gus mempromosikan kepercayaan yang lebih besar di kalangan klinik dan memaklumkan strategi yang lebih baik untuk pencegahan dan pengurusan strok di Banjarmasin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
AIS	Acute Ischemic Stroke
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under Curve
BLR	Bayesian Linear Regression
BMI	Body Mass Index
BN	Bayesian Network
CHS	Cardiovascular Health Study
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Tree
DTR	Decision Tree Regression
DNN	Deep Neural Network
EFS	Exhaustive Feature Selection
EVT	Endovascular Thrombectomy
FI	Features Importance
GA	Genetic Algorithm

GB	Gradient Boosting
GBM	Gradient Boosting Machine
HT	Hyperparameter Tuning
KNN	K-Nearest Neighbors
LaR	Lasso Regression
LoR	Logistic Regression
LDA	Linear Discriminant Analysis
LGBM	Light Gradient Boosting Machine
LOS	Length of Stay
LR	Linear Regression
MMR	Mixed-Mode Research
ML	Machine Learning
MNN	Model-based Neural Network
NB	Naïve Bayes
NLP	Natural Language Processing
NN	Neural Network
NNR	Neural Network Regression
PCA	Principal Component Analysis
PR	Polynomial Regression
REGARDS	Reasons for Geographic and Racial Differences in Stroke
RF	Random Forest
RFS	Recursive Feature Selection
RR	Ridge Regression
RIND	Reversible Ischemic Neurological Deficit
RQ	Research Question

SBFS	Step Backward Feature Selection
SFFS	Step Forward Feature Selection
SIE	Stroke in Evolution
SMOTE	Synthetic Minority Oversampling Techniques
SLR	Systematic Literature Review
SVM	Support Vector Machine
TB	Tree Boosting
TIA	Transient Ischemic Attack
XGBoost	Extreme Gradient Boosting
WHO	World Health Organization

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

1.1 Introduction

Stroke, a major global health issue, is defined by sudden neurological damage resulting from vascular disorders, leading to either an ischemic or hemorrhagic event (World Health Organization, 2018). Ischemic stroke occurs when a thrombus blocks a blood vessel in the cranial cavity, disrupting blood circulation and the crucial oxygen supply to cerebral tissue (Amarenco, 2020). This type of stroke is the most prevalent, representing 87% of all documented stroke events (Alexandrov, 2019).

Accurately identifying ischemic stroke is crucial for facilitating rapid treatment delivery (Agianto et al., 2022). This can significantly reduce the frequency and severity of strokes, improve patient prognosis, and lower healthcare costs (Wijaya et al., 2019). With precise classification models, early detection of stroke risk becomes possible,

paving the way for more effective intervention and prevention initiatives (Kumar et al., 2023).

However, developing such models requires extensive datasets and appropriate artificial intelligence methodologies to ensure reliable early detection outcomes (Lin et al., 2021). The ischemic stroke dataset encompasses various features that illustrate the associated risk factors, such as demographic statistics, medical history, and lifestyle choices (Clifford et al., 2019). Through careful analysis, researchers can identify patterns and correlations that enhance stroke risk prediction, improve early diagnostic capabilities, and develop more effective intervention strategies (Maida et al., 2020).

XGBoost, recognized for its excellent performance, stands out as a leading framework in machine learning, particularly beneficial for analyzing ischemic stroke pathology in the healthcare sector (Dhillon et al., 2021). Despite XGBoost's robust data analysis capabilities, a significant challenge often arises from the imbalance in the dataset (Fang & Deng, 2023). This imbalance can lead to suboptimal predictive accuracy, as the model tends to favor the more prevalent class within the dataset (Wang et al., 2021).

Therefore, utilizing various data balancing techniques, such as oversampling or undersampling, along with suitable evaluation methods, is crucial to ensuring that models provide accurate and reliable predictions of ischemic stroke (Hassan et al., 2022).

1.2 Research Motivation

The increase in stroke cases in Indonesia, particularly in Banjarmasin, underscores the urgent need for accurate prediction models and early intervention (Riset Kesehatan Dasar, 2020). For example, in South Kalimantan, the incidence of stroke surged from 739 cases in 2016 to 9,361 cases in 2018 across its 13 districts (Er Unja, 2022). Traditional risk assessment methods often fail to capture the complex nature of stroke risk, have limitations in big data integration, and depend on subjective judgment, which hinders rapid and appropriate diagnosis and treatment (Tseng & Noseworthy, 2021; Yu & Chen, 2023).

Although advances in machine learning (ML) have transformed data analysis in healthcare, applying it to ischemic stroke prediction still encounters significant complexities that previous studies have not adequately addressed (Yu & Chen, 2023). Key challenges include an extreme class imbalance in the dataset (where stroke cases are sporadic; for example, studies often report minority classes ranging from 5% to 8.6% of the initial data), which often leads to biased models and low accuracy in detecting critical cases (Kumar et al., 2023; Faura et al., 2021; Lin et al., 2021).

Additionally, many advanced ML models are often "black boxes," making it difficult for clinicians to understand the basis of predictions. In contrast, model interpretation is essential for actionable clinical decision-making and implementation in real care environments (Liu et al., 2019; Wang et al., 2021).

Consequently, this research aims to tackle these ongoing challenges, particularly with a dataset that captures the unique characteristics and severe imbalances of the population in Banjarmasin. While the XGBoost algorithm and the Synthetic Minority Over-sampling Technique (SMOTE) are established methodologies, the innovation of this study lies in the creation of a meticulously optimized modeling pipeline (Abiodun & Wreford, 2023; Li, 2024; Yang & Guan, 2022).

This pipeline tackles the significant complexity associated with the data while generating models that are both interpretable and relevant to clinical practices (Li, 2024). XGBoost was selected for its proven ability to manage complex datasets and produce accurate predictions, as well as its capacity to provide feature importance, which enhances model interpretation (Kaneko, 2023). SMOTE was essential in balancing the extreme class imbalance by creating synthetic samples for the minority class, ensuring the model can learn more effectively and accurately detect crucial minority classes (Kovács, 2019).

This study aims to produce an ischemic stroke prediction model that is not only accurate but also interpretable, reliable, and clinically applicable. It does so through systematic optimization using two specific hyperparameter tuning strategies and by exploring various data split ratios (70/30, 75/25, 80/20, 85/15). Importantly, a rigorous evaluation framework employing K-Fold Cross-Validation is utilized to report the mean and standard deviation, contributing to more effective prevention strategies and improved patient management in Banjarmasin.

1.3 Research Background

The number of stroke cases in Banjarmasin, Indonesia, has shown a concerning increase, with data from the Banjarmasin City Health Office recording approximately 8,739 stroke patient cases from 2018 to 2021 (Riset Kesehatan Dasar, 2020). This underscores the urgent need for accurate prediction models and early intervention. Traditional risk assessment methods often fail to capture the complex nature of stroke risk, have limitations in big data integration, and rely on subjective judgment, hindering rapid and appropriate diagnosis and treatment (Kumar et al., 2023; Yoshida et al., 2023).

Although advances in machine learning (ML) have transformed data analysis in healthcare, its application in ischemic stroke prediction continues to face significant complexities that previous studies have not adequately addressed (Alaka et al., 2020; Benzakoun et al., 2021). Algorithms like eXtreme Gradient Boosting (XGBoost) have emerged as promising options for stroke prediction, even outperforming models such as random forest and providing crucial feature interpretability (Fernandez-Lozano et al., 2021; He et al., 2022). Key challenges include a severe class imbalance in the dataset, where disease cases (the minority class) are considerably underrepresented, often making up only 5% of records belonging to the stroke class (Rana et al., 2021).

This imbalance can lead to biased models and diminish the accuracy of detecting crucial cases (Kumar et al., 2023; Harari et al., 2020). Additionally, optimizing the performance of advanced ML models like XGBoost requires systematic hyperparameter tuning (Ubaidillah et al., 2022). While this process is sufficient, it is

often not explored in depth or specifically tailored to fully realize the model's maximum potential when working with datasets having unique characteristics, such as local data (Grosser et al., 2020).

To tackle the prevalent issues of class imbalance and overfitting in medical datasets, this study employs the Synthetic Minority Over-sampling Technique (SMOTE). The research mainly concentrates on the XGBoost algorithm, comparing its performance with the Random Forest and Decision Tree algorithms as benchmarks. Random Forest serves as a fitting ensemble method for comparison with XGBoost due to their comparable complexity and robust performance, enabling a more equitable and informative evaluation than single models like Decision Tree (Vijiyakumar et al., 2019).

The Decision Tree was chosen for its clear interpretability, which enables clinicians to understand the model's underlying logic and its capacity to minimize the influence of outliers (Heldner et al., 2022). While a Decision Tree may not consistently achieve the highest accuracy, it frequently delivers competitive results. It is a reliable benchmark for evaluating the performance of ensemble methods like XGBoost (He et al., 2022).

This study specifically aims to evaluate the performance of the XGBoost algorithm by utilizing two distinct parameter tuning approaches and implementing the SMOTE technique. The primary goal is to enhance the accuracy of ischemic stroke prediction models and address limitations in previous studies regarding the management of extreme data imbalance and the lack of detailed, clinically actionable

model interpretations. Ultimately, this research seeks to develop a model that identifies and validates key risk factors through statistical testing, thereby contributing to more informed decision-making in stroke prevention and treatment.

1.4 Problem Statement

Predicting ischemic stroke in Banjarmasin presents significant challenges due to the inherent limitations of traditional predictive methodologies and the complexities associated with the data used. Conventional methods often fail to consider complex risk factors, cannot integrate big data, and depend too heavily on subjective medical assessments. This results in variations in diagnosis and treatment, which hinder early identification and effective intervention (Yu & Chen, 2023).

Moreover, significant challenges arise from the characteristics of the datasets used in stroke prediction. Many datasets, including those pertinent to the local context of Banjarmasin, often suffer from:

1. Extreme Class Imbalance: Ischemic stroke cases (the minority class) are significantly fewer than non-stroke cases (the majority class). This level of imbalance can be so severe that standard models, even when using basic balancing techniques, tend to be biased and struggle to detect the critically important minority class (Hassan et al., 2022). This underscores a fundamental failure in modeling high-impact medical data (Hasan & Hasan, 2020).

2. Missing and Inaccurate Data: Incomplete information can diminish the reliability of predictive models (Naar et al., 2020).

3. Inadequate Population Representation: Datasets that do not accurately reflect local demographic characteristics, like those in Banjarmasin, can lead to biased prediction outcomes (Hassan & Omar, 2021).

4. Non-linear Interactions Between Features: Traditional methods and some basic ML algorithms struggle to capture the genuinely complex and non-linear relationships among risk factors (Molnar et al., 2023).

5. Lack of Model Interpretability in Advanced Algorithms: While advanced ML models can achieve high accuracy, many operate as "black boxes," making it difficult for clinicians to grasp the reasoning behind predictions (Yang et al., 2022). This understanding is crucial for trust and adoption in a clinical setting (Yong & Gao, 2023).

Integrating traditional models with advanced techniques is often recommended as a promising strategy for enhancing predictive accuracy. However, this combination can lead to increased demands on computational resources and greater implementation complexity in clinical environments that require rapid decision-making, particularly in acute stroke management (Challen & Danon, 2023).

This research aims not only to compare models but also to provide innovative contributions in optimization and in-depth analytical methodologies to address the

limitations of existing predictive methods. It presents a more accurate, interpretable, and clinically relevant model for ischemic stroke prediction, particularly concerning the Banjarmasin population, while also showcasing the potential for generalization in future applications.

1.5 Research Question

The main emphasis of this study lies in the following research questions, which take into account the identified problem description and study objectives:

1. What are the most influential risk factors for ischemic stroke prediction identified through in-depth feature analysis of optimized XGBoost models using SHAP?

2. How can XGBoost-based machine learning models be optimized for predicting ischemic strokes on imbalanced datasets in Banjarmasin through hyperparameter tuning and rigorous validation?

3. How effectively does the systematic application of SMOTE enhance the accuracy of minority class detection in highly imbalanced ischemic stroke prediction data, and how can its interpreted contributions through advanced model explainers support more precise clinical decision-making in Banjarmasin?

1.6 Research Objectives

The following research question statement was the primary focus of this study, considering the study objectives:

1. To identify the most influential risk factors in predicting ischemic stroke through in-depth feature analysis using SHAP model explainers from the optimized XGBoost model, and to present novel insights that can guide targeted stroke prevention interventions in Banjarmasin.

2. To systematically develop an optimized ischemic stroke classification model using the XGBoost algorithm by implementing two specific hyperparameter tuning strategies and exploring various data split ratios within a rigorous validation framework, explicitly addressing the extreme class imbalance in datasets from the Banjarmasin population.

3. To assess the effectiveness of systematically applying the Synthetic Minority Over-sampling Technique (SMOTE) in improving ischemic stroke prediction model performance, particularly regarding minority class detection accuracy in highly imbalanced data conditions, and to examine SMOTE's interpretability and contributions to potential clinical decision-making through advanced model explainers in Banjarmasin.

1.7 Research Scope

Ischemic stroke remains a leading cause of morbidity and mortality worldwide, highlighting the critical need for public health discussions aimed at developing more effective prevention strategies. Research in this field emphasizes increasing awareness, identifying at-risk individuals, and advancing therapies. Banjarmasin, a region with a high prevalence of risk factors such as diabetes and hypertension, is a key focus for ischemic stroke research.

This investigation underscores Banjarmasin's significant vulnerability to cerebrovascular accidents, which have substantial public health implications. Therefore, evidence-based interventions are crucial for formulating effective programs and enhancing public awareness regarding the threat of ischemic stroke, thus facilitating risk mitigation.

The data collected for this study spans from 2012 to 2022 and includes 8,607 patients with twelve features. Of these, eleven features, adapted from Emon et al. (2020), are related to stroke risk factors and genetic factors as recommended by neurologists. This study employs machine learning techniques, specifically the XGBoost algorithm, and compares two types of parameters tuning based on previous studies.

To ensure methodological rigor and the novelty of the expected research, this study will also include Random Forest as an additional benchmark model that is more relevant for comparison with XGBoost (Ferdib-Al-Islam & Ghosh, 2021; Hasan &

Hasan, 2020). To ensure methodological rigor and the novelty of the expected research, this study will also include Random Forest as an additional benchmark model that is more relevant for comparison with XGBoost.

This study employs the Synthetic Minority Over-sampling Technique (SMOTE) with a data-splitting approach, using four ratios: 75/25, 70/30, 80/20, and 85/15. Methodological rigor in the validation process is enhanced by strictly adhering to a Train-Validation-Test split framework. All experiments will be conducted through multiple runs using K-Fold Cross-Validation to report results as means and standard deviations.

Additionally, this research will utilize statistical tests, including chi-square and ANOVA with the Tukey HSD test, to determine whether model results show significant differences between predictions and actual labels in categorical data. Model interpretability will be bolstered using advanced model explainers like SHAP to provide clinicians with more transparent and trustworthy insights.

1.8 The Importance of Research

While many risk assessments rely on statistical data to predict an individual's stroke risk, machine learning (ML) has shown remarkable potential in understanding non-linear interactions from large datasets, exceeding the abilities of traditional methods. Recent studies indicate that ML technology can profile patients and determine diagnoses and prognoses for at-risk individuals. However, current approaches often

concentrate solely on comparing the performance of forecasting techniques with fixed parameters or samples from the general population.

This research develops a risk prediction method that is not only based on statistical data; it specifically addresses the challenges of extreme class imbalance and improves model interpretability in clinical settings. Therefore, this study contributes new computer science and healthcare knowledge by utilizing a model designed to provide maximum and reliable results. This research aims to assist healthcare professionals and the government in mitigating the impact of stroke, thus allowing this study's planning and successful implementation processes to generate new knowledge.

1.8.1 Health and Government

In the context of health policy and government concerning stroke prediction, this research makes four significant contributions:

1. Promoting Evidence-Based Policy: Provides strong evidence to encourage the government to develop more effective policies for reducing stroke incidence and minimizing its socioeconomic impact.

2. Enhancing Accurate and Rapid Clinical Decision-Making: Equips medical professionals with more precise tools for diagnosing strokes, facilitating early treatment choices, and preventing serious complications. Interpretable models, supported by advanced model explainers, help clinicians understand the

rationale behind predictions, fostering trust and encouraging adoption in clinical settings.

3. Meeting Modern Medical Decision Support Needs: Delivers predictive, novel, and accurate decision support critically important in contemporary medical practice.

4. Developing Relevant Clinical Tools for Banjarmasin: Creates tools specifically usable in Banjarmasin's clinical settings to assist with stroke diagnosis and ensure management with a clear understanding of risks, accommodating local population characteristics and dominant risk factors identified in this study.

1.8.2 Computer Science (AI)

In computer science, particularly in the field of artificial intelligence, this research offers four significant contributions:

1. Accelerating Transformational Clinical Informatics Development: This initiative speeds up the development of transformational clinical informatics and ushers in a new era of machine learning utilization in healthcare decision-making.

2. Facilitating Further Research with Balanced and Interpretable Data: Makes it easier for other researchers to apply their models to datasets with more valid

results, especially regarding previously imbalanced data, thereby facilitating more accurate integrated clinical decision-making.

3. Simplifying Practical ML Applications for Medical Domains: Streamlines the application of integrated ML technology across various medical fields, including stroke, and develops practical ML applications using newer algorithms to achieve human-comparable accuracy.

4. Improving Computerized Stroke Care Systems: Enhances computerized stroke care systems to diagnose or predict prognostic trajectories more accurately, addressing data imbalances not optimally managed in previous studies while providing clear identification of important features through advanced model explainers.

1.9 Research Organization

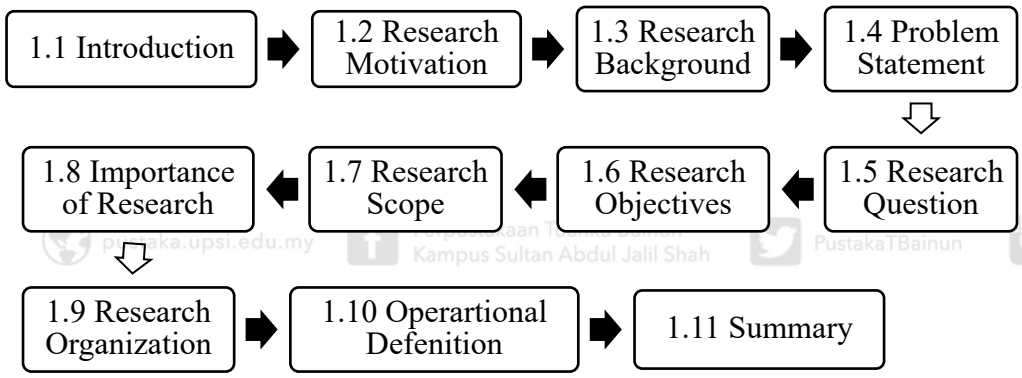
This section explained five chapters that serve as an overview while capturing the essence of this study. These chapters are arranged as follows:

The first chapter delves into the background of the particular study, including its problem statement, research objectives, questions, scope of the research, institutional affiliation, and operational definition, followed by a concluding abstract.

Figure 1.1 displays the research study structure, which comprises eleven sections, each elaborated on meticulously. **Section 1.1** offers the introduction, where all overarching concerns about the topic are provided alongside context and significance justification within the wider academic and practical context. **Section 1.2**, entitled "Research Motivation," articulates the rationale, including knowledge gaps and problems, that justifies conducting this specific piece of research.

Figure 1.1

Flow Diagram Introduction in Chapter 1



Section 1.3 includes a literature review outlining the relevant historical milestones and defining aspects of the theory that underpin and inform the research to situate it within an academic or practical setting appropriately.

The part (**Section 1.4**) clearly defines and succinctly highlights the specific concerns addressed by the research, thus further refining the scope of the study. The Research Questions (**Section 1.5**) derive from the Problem Statement and are articulated as specific, focused inquiries that will guide the research.

As seen in **Section 1.6**, objectives for any given study are its uniformly-significant outputs, which could involve determining certain relationships, validating defined hypotheses, or testing novel approaches aligned with posed questions.

Scope of research (**Section 1.7**) outlines delimitations of a given study, specifying which areas, populations, variables, or components would be included and which irrelevant ones would be excluded, hence adding clarity. Importance of Research (**Section 1.8**) contributes to how this research adds value to knowledge, practice, policy, or theory. In detailing sub-sections, such contributions are made into two main segments: health and governance, and computer science.

This can be found in **Section 1.9**, 'Organization of the Research', which describes the relationships between chapters and sections relative to their parts regarding framing construction, which is more than outlining.

To maintain coherence and clarity, the operational definitions (**Section 1.10**) delineate encounter critical terms detailing each definition with surgical precision, which are fundamental to the research, with sharp attention to detail.

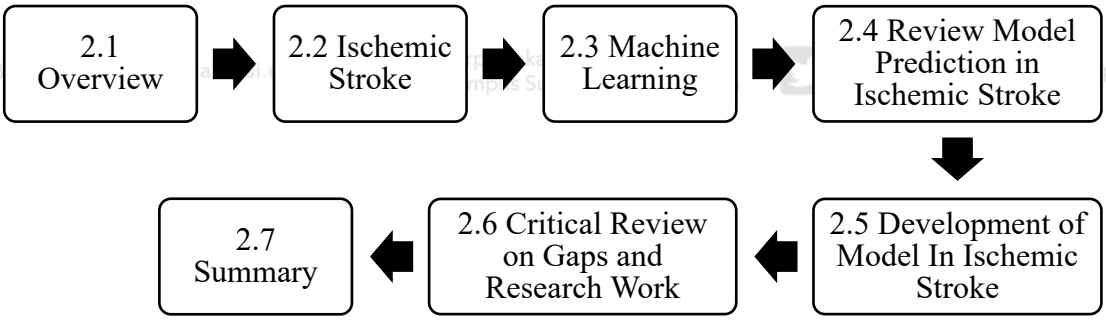
Additionally, in **Section 1.11**, the summary describes short prose capturing crucial ideas and concepts within the introduction, thus aiding smooth transition for readers into subsequent chapters. Combined with these eleven sections, this document provides structure while methodically and comprehensively addressing the research problem.

A literature review detailing all stroke classifications serves as the focus for chapter two. The literature review structure presented in **Figure 1.2** focuses on six detailed sections, which form part of an overarching discussion of applying machine learning techniques to predict ischemic strokes.

Section 2.1 presents background information introducing the focus of discourse concerning literature on predictive research related to ischemic strokes using advanced machine learning algorithms to provide context for the following discussion.

Figure 1.2

Literature Review Flow Diagram in Chapter 2



Section 2.2 presents an in-depth exploration of ischemic stroke, covering definitions, risk factors, signs and symptoms, impact, and relevant datasets. This section aims to deepen the understanding of the disease, emphasize its importance as a critical health issue, and highlight the need for more advanced predictive and diagnostic tools.

Section 2.3 introduces the concept of machine learning and discusses the problem of prediction models based on the literature review and the challenges faced in healthcare. **Section 2.4** reviews existing predictive models and methods for ischemic stroke, evaluating the performance of existing tools, algorithms, and approaches. It highlights the limitations and gaps in these models, including accuracy, data quality, and generalization issues.

Section 2.5 discusses the development of machine learning models explicitly aimed at ischemic stroke. This section explains aspects of using algorithms such as XGBoost, SMOTE, and Decision Tree as benchmarks for building effective prediction models.

Section 2.6 critically reviews the gaps in existing research, identifying limitations, unsolved problems, and deficiencies in current knowledge related to ischemic stroke and the application of machine learning models. This analysis forms the basis for further research and improving the accuracy and usability of existing models.

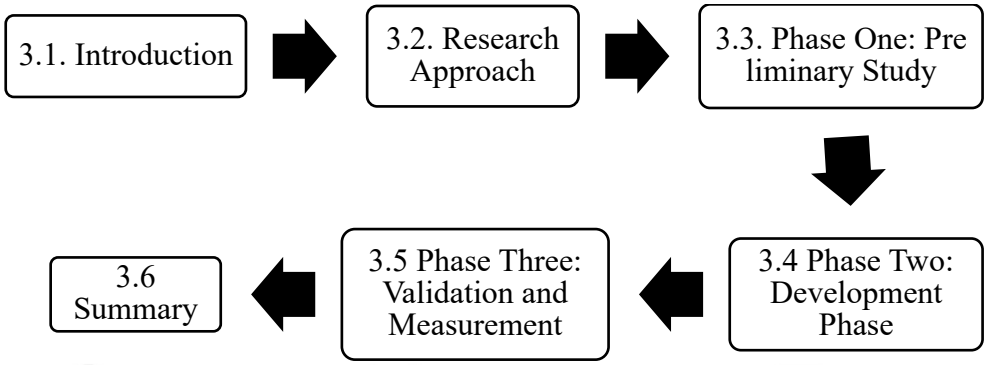
Finally, **Section 2.7** summarizes the key findings from the discussion, offers an overview of the topics covered, emphasizes the significance of machine learning in advancing ischemic stroke research, and outlines potential directions for future research.

Chapter 3 focuses on the classification of research methodology and research approach models. In **Figure 1.3**, the principal sections of **Chapter 3** are delineated,

each concentrating on a specific aspect of the research methodology. **Section 3.1** provides a comprehensive overview of **Chapter 3**, outlining the primary objectives of the chapter, the context of the research approach, and summarizing the structural organization of the chapter.

Figure 1.3

Methodology Flow Diagram in Chapter 3



Section 3.2 details the research approach employed in the study. This section specifies the type and methodology adopted, including experimental designs, data collection methods, and data analysis techniques. The approach is strategically crafted to fulfil the stated research objectives by aligning the research framework with the questions posed.

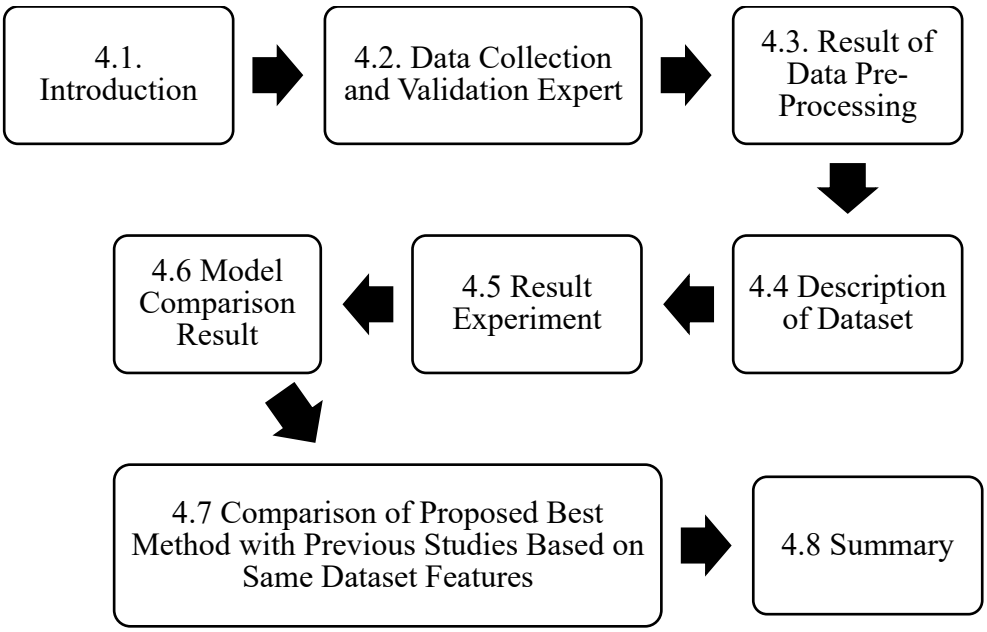
Section 3.3: The first research phase involves building a foundation by identifying problems, collecting data, and ensuring expert validation. This phase will answer the first part of the research question. In **Section 3.4**, the model development phase to predict ischemic stroke is described and enhanced by applying the SMOTE technique.

Section 3.5 elucidates the methods for measuring and validating the research outcomes. It addresses model evaluation by applying performance metrics such as accuracy, ROC/AUC, and confusion matrix. Furthermore, the validation process incorporates feature importance analysis, statistical tests, and a comparative analysis of model results against previous studies and benchmarks.

Finally, **Section 3.6** offers a summary of **Chapter 3** contents. This subsection revisits the main points discussed, including the research, development, and validation methodologies. The aim is to help the reader understand how the methods relate to the research objectives.

Figure 1.4

Findings Flow Diagram in Chapter 4



Chapter 4 assesses and presents the study's findings, data analysis, and experimental results. **Figure 1.4** illustrates the flow diagram of the results from the data analysis and experiments in **Chapter 4**. This diagram outlines the sequence of steps for each result based on the research methods detailed in **Chapter 3**. Below is an explanation of each component in the flowchart:

Section 4.1 introduces the overall structure and objectives of Chapter 4, providing an overview of the goals and steps involved in data collection, expert validation, data analysis, experimentation, and comparing experimental results with previous studies and the proposed method.

Section 4.2 details the methods for collecting and validating necessary data, employing professional expertise and techniques to ensure accuracy, reliability, and suitability for analysis. Subsequently, **Section 4.3** presents the outcomes of the data preprocessing phase, where raw dataset information undergoes cleaning, normalization, and transformation in preparation for further stages.

Section 4.4 explains the dataset utilized in the study, including its structure, features, size, and specific characteristics relevant to the research. It also analyzes the correlation of each feature with the target feature, stroke.

Section 4.2 describes the procedures for gathering and verifying the required data using professional expertise and techniques. This ensures that the data used in the study is accurate, reliable, and suitable for analysis. **Section 4.3** highlights the results

of the data preprocessing stage, where the raw data from the dataset is cleaned, normalized, and transformed to prepare it for subsequent steps.

Section 4.4 explains the dataset utilized in the study, including its structure, features, size, and specific characteristics relevant to the research. It also analyzes the correlation of each feature with the target feature, stroke.

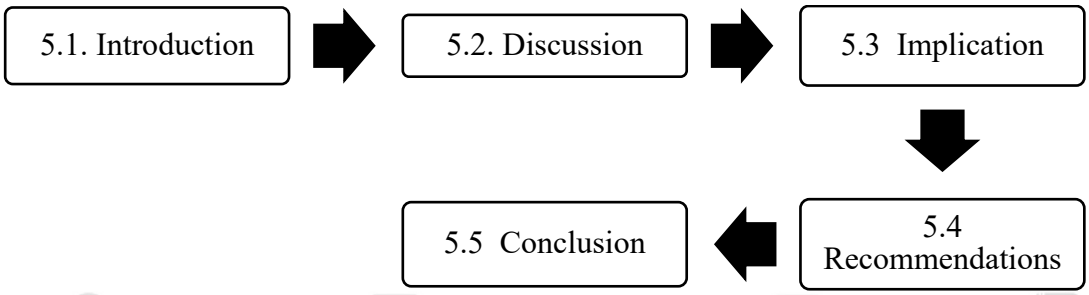
Section 4.5 presents the findings from the experimental analysis using the prepared data. This section details the experimental setup, methods, and performance metrics. For instance, it discusses how the data was divided into four ratios and the percentage increase in the dataset after applying SMOTE compared to the original data. Subsequently, the XGBoost algorithm and two types of hyperparameters are used, and a Decision Tree is evaluated both with and without SMOTE. Each experiment is categorized into four models based on the data-splitting ratio, yielding three results: the confusion matrix, statistical tests using Chi-Square, and feature importance.

Section 4.6 compares the results obtained from various models and techniques. This comparison helps assess the effectiveness of the proposed model against alternatives, addressing research questions one and three. **Section 4.7** reviews the current findings with previous studies with the same dataset. This section validates the improvements or differences achieved and compares prior research with the method proposed in this study.

Finally, **Section 4.8** concludes the chapter by summarizing the findings, insights, and implications, providing the reader with a comprehensive understanding of the chapter's content. This concluding section summarizes the discussed results and highlights the relationships among the various elements outlined earlier.

Figure 1.5

Discussion, Implication, Recommendations, and Conclusion Flow Diagram in Chapter 5



Chapter 5 Contains the discussion, implication, recommendations, and conclusion of the thesis. **Figure 1.5** presents the Introduction Flow Diagram in Chapter 5, outlining the structure and flow of topics to be discussed in this chapter. The following is an explanation of each component:

Section 5.1 provides a brief overview of Chapter 5, describing the scope of the discussion based on the results from the previous chapter. It then follows by describing the contributions derived from this research. It then addresses this study's limitations and concludes with a summary of the three preceding sections.

Section 5.2 discusses an in-depth analysis of the research's key findings, insights, or analyses. The results are interpreted comprehensively, linking them to existing knowledge or theoretical frameworks. **Section 5.3** outlines the implications encountered during the research process. This includes challenges, assumptions made, and areas where the research could be improved or expanded. **Section 5.4** recommendations for explaining future work on details. **Section 5.5** concludes the chapter by summarizing the key points, reflecting on the overall outcomes, and providing recommendations or implications for future research.

1.10 Operational Definition

This section delineates the significance of comprehending the terminologies employed within this study, as outlined below:

1. Dataset: A dataset is a systematically organized collection of data, typically formatted in a tabular structure, where individual observations are represented by rows and variables by columns. In medical research, datasets encompass patient records, diagnostic test results, and treatment outcomes, which are utilized to train machine learning models to predict health conditions or assess treatment efficacy (Liu et al., 2019).

2. EHR (Electronic Health Records): Electronic Health Records (EHRs) are digital representations of traditional paper-based patient records, which are crucial in storing comprehensive patient information. EHRs enhance the collection and

analysis of health data, enabling healthcare providers to make informed decisions based on a patient's medical history (Abedi et al., 2021).

3. Administrative Data: This term refers to data gathered primarily for administrative functions, such as billing and insurance claims. Administrative data are instrumental for healthcare research, yielding insights into service utilization, costs, and treatment outcomes (Matsui et al., 2022)
4. Preprocessing: *Preprocessing* is a critical phase that involves transforming raw data into a format more suitable for analysis and modelling. This process is crucial for improving data quality, directly affecting machine learning algorithms' performance. The importance of preprocessing cannot be overstated, as this process addresses various data-related issues, such as noise reduction, missing values, and data normalization, which can significantly impact the predictive performance of the model (Sari et al., 2021).
5. ML (Machine Learning): Machine learning constitutes a subset of artificial intelligence involving algorithms that empower computers to learn from data and make predictions. In stroke detection, machine learning can discern patterns within medical or imaging data to facilitate early detection of strokes, which is critical for effective treatment (Chen et al., 2020).
6. SMOTE: The Synthetic Minority Over-Sampling Technique (SMOTE) is a statistical method that addresses class imbalance within a dataset by generating synthetic samples from minority classes. This technique is particularly pertinent in medical datasets where certain conditions, such as strokes, are infrequent,

thereby mitigating model bias towards the majority class (Ferdib-Al-Islam & Ghosh, 2021).

7. XGBoost: XGBoost is a highly efficient implementation of the gradient boosting framework widely utilized for classification and regression tasks. Its prominence in medical data analysis can be attributed to its capacity to manage large datasets and enhance prediction accuracy, rendering it suitable for applications in disease diagnosis (Chen et al., 2017).

8. Features Importance: Feature importance is a fundamental concept in machine learning, especially in algorithms such as Extreme Gradient Boosting (XGBoost). These algorithms' ability to assign importance values to individual features allows practitioners to identify which variables exert the greatest influence on model predictions. This functionality improves the model's interpretability and serves as an important guide in the feature selection process, thus optimizing the overall modelling framework (Kaneko, 2023; Molnar et al., 2023).

9. Decision Tree: A decision tree is a predictive model that translates observations into inferences regarding a target value. This model is compelling in health classification due to its interpretability and support for clinical decision-making processes (Inoue et al., 2020).

10. Hyperparameter Tuning: This process entails optimizing the parameters of a machine learning model to enhance its performance. In medical applications,

hyperparameter tuning can significantly influence prediction accuracy, ensuring the model is well-suited to adapt to novel data (Ubaidillah et al., 2022).

11. Confusion Matrix: A confusion matrix is a tabular representation utilized to assess the performance of a classification model. This matrix summarizes correct and incorrect predictions, offering insights into the model's accuracy, sensitivity, and specificity, which are critical for evaluating diagnostic tests (Ting, 2017).

12. Split Data: This concept entails partitioning a dataset into training and testing subsets. The training data is employed to develop the model. In contrast, the testing data is utilized to evaluate the model's performance, a practice essential for ensuring the model's generalizability to new data (Kahloot & Ekler, 2021).

13. Training Data: Training data constitutes the dataset segment used to train the machine learning model. This data enables the model to learn from existing patterns, particularly in medical diagnostics, where precise predictions can significantly impact patient outcomes (Kahloot & Ekler, 2021).

14. Testing Data: Testing data is employed to evaluate the performance of the trained model. This assessment is crucial for determining the model's predictive capabilities on new data, ensuring reliability in clinical diagnostics (Kahloot & Ekler, 2021).

15. Accuracy: Accuracy quantifies the frequency with which the model generates correct predictions. In medical diagnostics, high accuracy is paramount to minimize misdiagnosis and safeguard patient safety (Ting, 2017).

16. Sensitivities: Sensitivity, also known as the valid positive rate, measures the proportion of actual positive results that the model accurately identifies. High sensitivity is essential for the early detection of diseases, such as strokes, where prompt intervention can be life-saving (Ting, 2017).

17. Specificity: Specificity gauges the proportion of actual negative results that the model correctly identifies. This metric is crucial in medical testing to ensure that healthy individuals are not misdiagnosed, preventing unnecessary anxiety and treatment (Ting, 2017).

18. ROC/AUC: The Receiver Operating Characteristic (ROC) curve plots the valid positive rate against the false positive rate to represent the model's diagnostic capabilities at various thresholds. The Area Under the Curve (AUC) measures the model's overall performance, with higher AUC values indicating better diagnostic capabilities. This metric is particularly relevant in evaluating the effectiveness of medical tests (Ting, 2017).

19. Validation Set: A subset of data utilized for model evaluation during the development and hyperparameter tuning phase, distinct from the training subset and the final test set. This aids in optimal model selection and mitigates overfitting on the test set (Purba et al., 2022).

20. K-Fold Cross-Validation: A model evaluation technique that divides the dataset into 'K' folds. The model is trained 'K' times, each time using 'K-1' folds for training and a different fold for validation. Performance results are then averaged across all 'K' iterations to provide a more robust and stable estimate of the model's capability. This addresses the limitations of a single random split (Wen et al., 2017).

21. Mean and Standard Deviation: Statistical measures used to report performance metrics from multiple runs. The mean indicates the average performance of the model, while the standard deviation reflects the variability or stability of that model's performance across different iterations (Rueangket et al., 2022).

22. SHAP (SHapley Additive exPlanations) is a model interpretation method grounded in cooperative game theory. It provides Shapley values for each feature, representing that feature's average and equitable contribution to the model's prediction. This helps clarify "why" the model made a specific prediction, both globally and for individual cases (Fang & Deng, 2023).

23. Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the class that represents the mode of the individual trees (classification) or the mean/average prediction (regression). It is known for its accuracy and its better ability to manage overfitting compared to single decision trees, making it a relevant benchmark for boosting models like XGBoost (Saragih et al., 2020).

Table 1.1

Feature Definition

No.	Feature	Definition
1	Age	This feature indicates the age of a person. It is numerical data.
2	Gender	This feature indicates a person's gender. It's categorical data.
3	Hypertension	This characteristic indicates whether this person has hypertension or not. This is numerical data.
4	Genetic factors	This trait is characterized by a stroke history that is associated with recurrent strokes.
5	Work type	This function represents the person's work scenario. This is categorical data.
6	Residence Type	A residence type refers to the classification of a dwelling place based on various factors such as ownership, purpose, and location. This is categorical data.
7	Heart disease	This attribute indicates whether or not this person has heart disease. This is numerical data.
8	Avg. glucose level	This attribute means what the person's glucose level was. This is numerical data.
9	Ever married	This feature represents a person’s married status. It is a categorical data.
10	Smoking status	This attribute is related to the person's smoking status. This is categorical data.
11	Stroke	This function indicates whether or not the person has had a stroke in the past. This is numerical data. Here, each attribute row is a decision class, and the remaining attribute is a response class.

Table 1.1 delineates a comprehensive array of demographic and health characteristics pertinent to the analysis of stroke risk factors. The " Age " feature is a numeric variable that quantifies the individual's age, whereas the "Gender" feature is categorical, classifying the respondent's gender.

The table also incorporates "Hypertension" and "Heart Disease" as numeric attributes, indicating the presence or absence of these medical conditions. The "Genetic Factors" feature elucidates the genetically inherited predisposition associated with a history of recurrent strokes.

Moreover, the employment and living conditions of individuals are characterized by the "Type of Employment" and "Type of Residence" features, both of which are categorical variables that describe the nature of employment and the individual's living environment.

The features "Average Glucose Level" and "Body Mass Index (BMI)" are numerical variables that represent an individual's average glucose level and body mass index, respectively. Furthermore, the categorical variables "Ever Married" and "Smoking Status" provide insights into an individual's marital status and smoking behaviours.

The final feature, "Stroke," is a numeric variable that indicates whether the individual has experienced a stroke. This collection of features enables the exploration of potential correlations between demographic, medical, and lifestyle factors and the risk of stroke, with "Stroke" serving as the response variable. The remaining features

function as predictor variables or decision classes that aim to predict or elucidate this outcome variable.

This dataset has been adapted from the research conducted by Emon et al. (2020), which assessed the influence of these features on stroke outcomes. The inclusion of the "Genetic Factors" feature was informed by recommendations from health experts, particularly within the context of ischemic stroke management.

1.11 Summary

This chapter highlights ischemic stroke as a significant global health issue, accounting for approximately 87% of all stroke cases due to blocked blood vessels in the brain, which leads to tissue damage from oxygen deprivation. The research emphasizes the importance of understanding non-modifiable risk factors, such as age and family history, along with modifiable factors, including hypertension and smoking, as these contribute to long-term disability and economic burden, particularly in low and middle-income countries.

This study addresses dataset imbalance by employing machine learning techniques, notably Synthetic Minority Over-sampling Technique (SMOTE) and the XGBoost algorithm, to enhance prediction accuracy. Importantly, it utilizes various data split ratios and hyperparameter tuning within a rigorous validation framework, including multiple runs with mean and standard deviation reporting, to

comprehensively evaluate model performance. Random Forest has been incorporated as a more relevant benchmark model for fair comparison.

The primary objectives of this study are to identify key risk factors, develop effective and interpretable predictive models using advanced model explainers like SHAP, and facilitate early identification of high-risk individuals. Therefore, this study is expected to reduce stroke incidence and healthcare costs, promote evidence-based decision-making in health policy, and foster innovative approaches for stroke prevention and management.