THE PROGNOSTIC AND PREDICTIVE MODELLING OF MORTALITY AMONG ACUTE STROKE PATIENTS IN PENINSULAR MALAYSIA

DR. CHE MUHAMMAD NUR HIDAYAT BIN CHE NAWI

UNIVERSITI SAINS MALAYSIA

2024

THE PROGNOSTIC AND PREDICTIVE MODELLING OF MORTALITY AMONG ACUTE STROKE PATIENTS IN PENINSULAR MALAYSIA

by

DR. CHE MUHAMMAD NUR HIDAYAT BIN CHE NAWI

Dissertation submitted in partial fulfilment of the requirements for the degree of
Doctor of Public Health
(Epidemiology)

MARCH 2024

ACKNOWLEDGEMENTS

First and foremost, I wish to dedicate this endeavour to Allah S.W.T., the Almighty, for the wisdom and resources bestowed upon me, enabling the successful completion of this project. I extend my heartfelt gratitude to the individuals who supported me throughout this journey.

To my beloved wife, Nurul Fatihah binti Mohamad Nasir, and my three children, Che Fatimah Azzahrah, Che Muhammad Noah Arman, and Che Khadeeja Aafiya, I want to express my deepest appreciation for your unwavering understanding and support during this challenging journey. Despite facing numerous health-related difficulties and the need to divide our attention, I am confident that these trials have made us stronger than ever before.

I am also immensely thankful to the lecturers of the Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia, for imparting essential public health research skills. My sincere gratitude extends to my supervisor, Professor Dr. Kamarul Imran Musa, and co-supervisor, Dr. Suhaily Mohd Hairon, as well as Associate Professor Dr. Wan Nur Nafisah Wan Yahya from Universiti Kebangsaan Malaysia. Their consistent guidance, input, and extensive knowledge have been instrumental in successfully completing this demanding Doctor of Public Health research project.

I extend my appreciation to the members of my research team (as listed in Appendix A), whose contributions were invaluable at every phase of the study. Additionally, I cannot thank the medical record staff, officers, and directors of the participating hospitals—Hospital Universiti Sains Malaysia, Hospital Canselor Tuanku Muhriz Universiti Kebangsaan Malaysia, Hospital Seberang Jaya, Hospital Raja Perempuan

Zainab II, and Hospital Sultanah Nur Zahirah—enough for their invaluable assistance in facilitating the data collection process for this research.

To my dear colleagues and others who indirectly contributed to assembling these pieces, may Allah bless each one of you. Lastly, I would like to express my gratitude to the Ministry of Health, Malaysia, for providing me with the opportunity and sponsorship to enhance my competency in public health.

DECLARATION

I declare that the thesis has been composed by myself in the manuscript-based thesis writing as an alternative format approved by the School of Medical Sciences, Universiti Sains Malaysia, and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated in the list of manuscripts. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Dr. Che Muhammad Nur Hidayat bin Che Nawi

Candidate ID: P-UD0008/20

Date: 2 March 2024

TABLE OF CONTENTS

LIST OF PAPER

This Doctor of Public Health (DrPH) dissertation contains four papers:

1. Che Muhammad Nur Hidayat Che Nawi, Suhaily Mohd Hairon, Wan Nur Nafisah Wan Yahya, Wan Asyraf Wan Zaidi, Mohd Rohaizat Hassan, Kamarul Imran Musa. *Machine Learning Application: A Bibliometric Analysis from a Half-Century of Research on Stroke*. (This paper was published in Cureus, DOI: 10.7759/cureus.44142. Please refer Appendix B).
2. Che Muhammad Nur Hidayat Che Nawi, Suhaily Mohd Hairon, Wan Nur Nafisah Wan Yahya, Wan Asyraf Wan Zaidi, Mohd Rohaizat Hassan, Kamarul Imran Musa. *Stroke Mortality Prediction in Malaysia: Evaluating Survival Rates and Comparing Glasgow Coma Scale and NIH Stroke Scale– A Multicentre Analysis*. (This manuscript draft will be submitted to International Medical Journal Malaysia for publication)
3. Che Muhammad Nur Hidayat Che Nawi, Suhaily Mohd Hairon, Wan Nur Nafisah Wan Yahya, Wan Asyraf Wan Zaidi, Mohd Rohaizat Hassan, Kamarul Imran Musa. *Machine Learning Models for Predicting Stroke Mortality in Malaysia: An Application and Comparative Analysis*. (This paper was published in Cureus, DOI: 10.7759/cureus.50426. Please refer Appendix C).
4. Che Muhammad Nur Hidayat Che Nawi, Suhaily Mohd Hairon, Wan Nur Nafisah Wan Yahya, Wan Asyraf Wan Zaidi, Mohd Rohaizat Hassan, Kamarul Imran Musa. *A Novel Machine Learning Approach for Predicting Stroke Mortality among Acute Stroke Patients in Malaysia*. (This manuscript draft will be submitted to International Medical Journal Malaysia for publication).

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF FORMULA

LIST OF ABBREVIATIONS

MIST	Malaysian Ischemic Stroke Mortality Prediction Tool
NIHSS	National Institute of Health Stroke Score
AIC	Akaike Information Criterion
AUC	Area Under the Curve
S-CVI/Ave	Scale-Level Content Validity Index Averaging Method
S-FVI/Ave	Scale-Level Face Validity Index Averaging Method
DALYs	Disability-Adjusted Life Years
LMICs	Low/Middle-Income Countries
THRIVE	Totaled Health Risks in Vascular Events
ISCORE	Ischemic Stroke Predictive Risk Score
ML	Machine Learning
SVM	Support Vector Machine
RSF	Random Survival Forest
ANN	Artificial Neural Network
DT	Decision Tree
RF	Random Forest
GBD	Global Burden of Disease
CVD	Cardiovascular Disease
TRIPOD	Transparent Reporting of A Multivariable Prediction Model For Individual Prognosis or Diagnosis
mRS	Modified Rankin Scale
GWTG	Get With The Guideline

ESRS	Essen Stroke Risk Score
SOAR	Stroke Subtype, Oxford Community Stroke Project Classification, Age, Pre-Stroke Modified Rankin
PLAN	Preadmission Comorbidities, Level of Consciousness, Age, And Neurologic Deficit
QRISK	Cardiovascular Disease Risk Algorithm
EURO-SCORE	European System For Cardiac Operative Risk Evaluation
HTN	Hypertension
DM	Diabetes Mellitus
AF	Atrial Fibrillation
MI	Myocardial Infarction
TIA	Transient Ischaemic Attack
OCSP	Oxfordshire Community Stroke Project
MERCI	Mechanical Embolus Removal in Cerebral Ischemia
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
Cox-EN	Cox Model With Elastic Net
EN	Elastic Net
CPH	Cox Proportional Hazards
LASSO	Least Absolute Shrinkage and Selection Operator

AIC	Akaike Information Criterion
C-index	Concordance Index
D-index	Discriminative Index
RCSN	Registry of The Canadian Stroke Network
eHealth	Electronic Health Technology
GCS	Glasgow Coma Scale
IV	Intravenous
CKD	Chronic Kidney Disease
HPL	Hyperlipidaemia
H-index	Hirsch Index
HUSM	Hospital Universiti Sains Malaysia
HCTMUKM	Hospital Canselor Tuanku Muhriz Universiti Kebangsaan Malaysia
HRPZII	Hospital Raja Perempuan Zainab Ii
HSJ	Hospital Seberang Jaya
HSNZ	Hospital Sultanah Nur Zahirah
ICD-10	International Classification of Diseases, 10th Revision
NPRS	National Population Registration System
SD	Standard Deviation
Adj. HR	Adjusted Hazard Ratio
CI	Confidence Interval
SD	Standard Deviation
SE	Standard Error

DF	Degree Of Freedom
ROC Curves	Receiver Operating Characteristic (ROC) Curves
HTML	Hyper Text Markup Language
URL	Uniform Resource Locator
I-CVI	Item-Content Validity Index
I-FVI	Item-Face-Validity-Index
JEPeM	Jawatankuasa Etika Penyelidikan Manusia
NMRR	National Medical Research Register
TIPPS	Tabung Insentif Pembangunan Pengajian Siswazah
SCP	Single Country Production
MCP	Multiple Country Production
COVID-19	Coronavirus Disease Of 2019
USA	United State Of America
STAIR	Stroke Therapy Academic Industry Roundtable
STEPS	Stem Cell Therapies as An Emerging Paradigm In Stroke
WSO	World Stroke Organization
SITS	Safe Implementation of Treatments In Stroke
VISTA	Virtual International Stroke Trials Archive
WHO	World Health Organization
SDLC	Software Development Life Cycle
ChatGPT	Chat Generative Pre-Trained Transformer

LIST OF SYMBOLS

>	More than
<	Less than
=	Equal to
≥	More than and equal to
≤	Less than and equal to
%	Percentage
+	Plus

ABSTRAK

PROGNOSTIK DAN PEMODELAN RAMALAN KEMATIAN DALAM KALANGAN PESAKIT STROK AKUT DI SEMENANJUNG MALAYSIA

Latar belakang: Evolusi teknologi digital dan kecerdasan buatan yang pesat telah mempengaruhi perubahan aplikasi pembelajaran mesin dalam pemodelan ramalan bagi penyakit strok dan akibatnya. Peningkatan morbiditi dan mortaliti akibat strok di Malaysia, cabaran dalam pengurusan klinikal dan prognostik menyebabkan terdapat keperluan yang mendesak untuk melakukan pemodelan yang tepat dalam meramalkan prognosis penyakit ini.

Objektif: Kajian ini bertujuan untuk menganalisis trend dalam penerbitan yang berkaitan dengan aplikasi pembelajaran mesin dalam pemodelan strok dan akibatnya, mengenal pasti faktor prognostik, melakukan pemodelan ramalan untuk kematian dalam kalangan pesakit strok akut, dan membangunkan aplikasi berasaskan laman sesawang untuk ramalan kematian strok akut dengan menggunakan data daripada pelbagai pusat strok di Malaysia merangkumi tahun 2016 hingga 2021.

Metodologi: Pendekatan kajian ini merangkumi pelbagai aspek: bermula dengan analisis bibliometrik menggunakan data Scopus dan Web of Science serta diikuti dengan analisis kohort retrospektif seramai 950 pesakit strok akut di lima buah hospital di Semenanjung Malaysia. Analisis Kemandirian *(Survival Analysis)* dan pelbagai teknik pemodelan ramalan, termasuk regresi Cox, *Support Vector Machine* (SVM), dan *Random Survival Forest* (RSF) digunakan dalam kajian ini. *Malaysian Ischemic Stroke Mortality Prediction Tool* (MIST) pula dibangunkan dengan melibatkan proses kebolehpercayaan dan kesahan instrument bersama pakar dan pengguna aplikasi ini.

Hasil dapatan: Analisis bibliometrik menunjukkan trend yang tinggi dalam bidang penyelidikan pembelajaran mesin bagi penyakit strok dengan kerjasama global yang signifikan. Kajian retrospektif pula mendapati purata umur permulaan menghidap strok adalah pada usia 63.15 (13.09) tahun, lelaki (n=552, 58.1%) dan etnik Melayu (n=771, 81.7%) mendominasi dengan ketepatan ramalan yang lebih tinggi bagi skala *National Institute of Health Stroke Score (NIHSS)* (nilai signifikan statistik yang lebih tinggi, nilai *Akaike Information Criterion (AIC)* yang lebih rendah, Indeks C yang lebih tinggi, dan trend penurunan yang beransur-ansur bagi keluk *survival Kaplan-Meier*) berbanding *Glasgow Coma Scale (GCS)* untuk kematian berkaitan strok. Pemodelan menggunakan SVM menunjukkan ketepatan ramalan yang lebih baik, dibuktikan oleh nilai *Area Under the Curve* (AUC) pada waktu 3 bulan, 1 tahun dan 3 tahun 0.842, 0.846, dan 0.791, dengan indeks D 5.31 (95% CI: 3.86, 7.30), indeks C 0.803 (95% CI: 0.758, 0.847), dan skor Brier antara 0.103 hingga 0.220. Para pakar dan pengguna MIST, mengakui bahawa aplikasi ini dapat memberikan ketepatan ramalan yang tinggi dan mesra pengguna dengan nilai kebolehpercayaan dan kesahan instrumen, *Scale-Level Content Validity Index (S-CVI/Ave)* 0.99 *dan Scale-Level Face Validity Index (S-FVI/Ave)* sebanyak 0.98.

Kesimpulan: Teknik pembelajaran mesin tidak asing lagi dalam penyelidikan strok, dan kian mendapat sambutan dan perhatian global serta sains perkomputeran. Kajian ini menekankan keperluan bagi Malaysia untuk mempunyai pemodelan ramalan yang efektif bagi penyakit strok, dengan SVM menunjukkan kemampuan yang tinggi dalam pemodelan ramalan kematian bagi penyakit ini. MIST, aplikasi atas talian yang telah disahkan, akan menjadi pemangkin yang signifikan dalam meningkatkan pengurusan penyakit strok melalui pemodelan ramalan kematian yang tepat.

Kata kunci: pemodelan prognostik, pemodelan ramalan, strok akut, mortaliti, Malaysia

ABSTRACT

THE PROGNOSTIC AND PREDICTIVE MODELLING OF MORTALITY AMONG ACUTE STROKE PATIENTS IN PENINSULAR MALAYSIA

Background: The rapid evolution of digital technology and artificial intelligence has revolutionized the application of machine learning in predicting stroke outcomes. The increasing burden of stroke in Malaysia, characterized by its impact on mortality and morbidity, underscores the need for accurate mortality prediction models. This need is heightened by the challenges in clinical decision-making and prognostic management, driving the development of various prognostic models and tools.

Objective: This study aimed to analyse trends in publications related to the application of machine learning in stroke outcome modelling, identify prognostic factors, perform predictive modelling for mortality among acute stroke patients, and develop a web-based application for stroke mortality prediction using data sourced from multiple stroke centres in Malaysia spanning the years 2016 to 2021.

Methodology: Our methodology spans a multifaceted approach: starting with a bibliometric analysis using Scopus and Web of Science data, followed by a retrospective cohort analysis of 950 stroke patients across five hospitals in peninsular Malaysia. We utilized survival analyses and an array of predictive modelling techniques, including Cox regression, Support Vector Machine (SVM), and Random Survival Forest (RSF). The development of Malaysian Ischemic Stroke Mortality Prediction Tool (MIST) involved a rigorous process of content and face validation with domain experts and users.

Results: The bibliometric analysis delineated a robust trend in machine learning research in the realm of stroke, punctuated by significant global collaborations. The retrospective study revealed a mean stroke onset age of 63.15 (13.09) years, with a male (n=552, 58.1%) and Malay ethnicity (n=771, 81.7%) predominance and a higher predictive precision of the National Institute of Health Stroke Score (NIHSS) scale (higher statistical significance, lower Akaike Information Criterion (AIC) values, a higher C-Index, and a more gradual decline in Kaplan-Meier survival curves) over Glasgow Coma Scale (GCS) for stroke-related mortality. Notably, the SVM model demonstrated superior predictive accuracy, evidenced by 3-month, 1-year, and 3-year time-dependent Area Under the Curve (AUC) values of 0.842, 0.846, and 0.791, a D-index of 5.31 (95% CI: 3.86, 7.30), a C-index of 0.803 (95% CI: 0.758, 0.847), and Brier scores ranging from 0.103 to 0.220. MIST, following comprehensive validation, was highly acclaimed by experts and users for its predictive accuracy and user-friendliness with Scale-Level Content Validity Index (S-CVI/Ave) and Scale-Level Face Validity Index (S-FVI/Ave) of ≥ 0.99 and ≥ 0.98, respectively.

Conclusion: Machine learning techniques are increasingly adopted in stroke research, facilitated by global collaborations and advancements in computational science. The study's findings highlight the need for effective predictive models in Malaysia, with SVM showing superior performance in mortality prediction. MIST, as a validated online tool, offers significant potential for enhancing stroke care and public health through accurate mortality risk estimation.

Keywords: prognostic modelling, predictive modelling, acute stroke, mortality, Malaysia

CHAPTER 1

INTRODUCTION

1.1 Overview of stroke

Stroke is clinically defined as the 'rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer or leading to death, with no apparent cause other than of vascular origin' (Aho et al., 1980). In a broader sense, strokes can be categorized based on their pathophysiology, including ischemic stroke, primary intracerebral haemorrhage, subarachnoid haemorrhage, and strokes with undetermined aetiology (Valery L. Feigin et al., 2003).

Over the past three decades, the challenge posed by stroke has remained persistent, marked by increasing rates of both mortality and disability. Globally, between 1990 and 2019, there was a significant upsurge in various stroke-related metrics, including a 70.0 percent increase in stroke incidents, a 43.0 percent rise in mortality rates, a 102.0 percent surge in prevalence, and a substantial 143.0 percent elevation in disability-adjusted life-years lost (DALYs). Notably, this growing burden of stroke is predominantly concentrated in Low- and Middle-Income Countries (LMICs), accounting for 86.0 percent of stroke-related fatalities and an alarming 143.0 percent of DALYs (Valery L. Feigin et al., 2021).

In Malaysia, much like many other nations, stroke remains a leading cause of mortality and a significant contributor to long-term disability(Tan and Venketasubramanian, 2022). In the year 2019, Malaysia recorded a staggering 47,911 incident cases, 19,928 fatalities, 443,995 prevalent cases, and a loss of 512,726 Disability-Adjusted Life Years (DALYs) attributed to stroke (Valery L. Feigin et al., 2021; Tan and Venketasubramanian, 2022).

This grim reality underscores the pressing need for improved methods that can inform treatment decisions and enhance the prognostic outlook for stroke patients. A crucial strategy in this regard is the prediction of long-term outcomes, particularly stroke mortality, among individuals experiencing acute stroke episodes. Prediction models, which amalgamate patient data and healthcare procedures to anticipate specific future health events, such as stroke mortality, play a pivotal role in stroke prevention (Fahey et al., 2018).

1.2 Modelling of stroke outcome

In the realm of disease modelling, two approaches have garnered significant attention: model-based prediction and model-free prediction. Both prediction models draw upon patient characteristics, clinical features, and care processes to estimate the likelihood of experiencing a specific future outcome, such as mortality (as depicted in **Figure 1.1**). These models have proven valuable in the context of primary stroke prevention by enabling healthcare professionals to accurately predict mortality risk (Collins et al., 2015). This predictive capability empowers clinicians to effectively communicate prognosis to patients and their families, facilitating informed decision-making regarding treatment strategies and care plans. Model-based prediction encompasses methods like logistic regression and Cox proportional hazard regression models. Conversely, model-free prediction employs machine learning techniques to make predictions (Fahey et al., 2018; W. Wang et al., 2020).

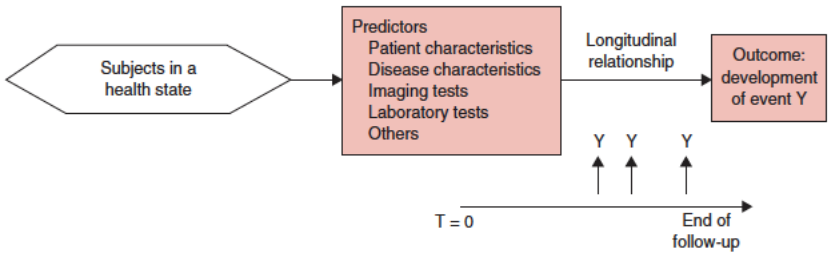

Figure 1.1 Schematic representation of prognostic modelling (Collins et al., 2015)

Within the domain of stroke research, model-based prediction has been exemplified by the use of tools like the Ischemic Stroke Predictive Risk Score (ISCORE) and the Totaled Health Risks in Vascular Events (THRIVE) score. These tools have not only influenced health service planning and policymaking but have also aided in clinical decision-making, diagnostic work-up, and therapy selection for high-risk groups (Flint et al., 2015; Saposnik et al., 2011).

The advent of increased computational power and the availability of model-free prediction methods, especially machine learning (ML), have generated considerable interest. ML leverages extensive, routinely collected datasets to provide reliable, personalized prognoses for stroke outcomes. Among the ML models employed for stroke mortality prediction are artificial neural networks (ANN), Naïve Bayes, support vector machines (SVM), decision trees (DT), and random forests (RF) (W. Wang et al., 2020).

1.3 Problem statement and study rationale

1.3.1 Problem statement

The field of stroke research represents a critical area in healthcare, given the substantial burden of stroke-related morbidity and mortality. Despite significant advances in medical research, several critical gaps in the existing body of knowledge related to stroke outcomes, particularly in Malaysia, have yet to be adequately addressed.

Firstly, to our knowledge, no previous study has undertaken a comprehensive bibliometric analysis to systematically explore the trends and patterns in the application of machine learning techniques within the context of stroke research. This gap in the literature hinders our understanding of the evolution of machine learning applications in this domain, impeding the identification of potential research directions and areas for improvement.

Secondly, there remains a dearth of studies focusing on stroke survival in Malaysia. While some studies have examined stroke outcomes in the Malaysian population, the majority have relied on descriptive analyses, with only a limited number utilizing the Cox proportional hazards model (Kooi Cheah et al., 2016). This scarcity of comprehensive survival analyses impedes the development of tailored interventions and policy recommendations aimed at improving stroke survival rates within the country.

Furthermore, the utilization of machine learning for stroke outcome prediction, specifically mortality prediction, in the Malaysian context remains underexplored. The potential of machine learning algorithms to provide accurate, individualized prognoses represents a promising avenue for enhancing stroke management and patient care (Fahey et al., 2018; W. Wang et al., 2020). However, the absence of research in this area deprives healthcare practitioners of valuable tools to inform clinical decisions and resource allocation.

Finally, despite the growing interest in machine learning for healthcare applications, there is currently no prediction tool available for stroke mortality specific to the Malaysian population, utilizing machine learning as a backend algorithm. Such a tool would not only contribute to the advancement of stroke research but also address

a pressing need in the Malaysian healthcare landscape by providing a customized and data-driven approach to predicting stroke outcomes.

1.3.2 Study rationale

The rationale for this research study is multifaceted and driven by the critical gaps identified in stroke research within the Malaysian context. Firstly, conducting a comprehensive bibliometric analysis of the application of machine learning in stroke research will provide a foundational understanding of the current state of the field. This analysis will help identify emerging trends, key contributors, and potential areas for further exploration, ultimately guiding future research endeavours and resource allocation.

Secondly, investigating stroke survival in Malaysia is of paramount importance, considering the nation's unique sociodemographic characteristics and healthcare infrastructure. By employing advanced statistical techniques, including machine learning, we aim to gain deeper insights into the factors influencing stroke survival, facilitating the development of targeted interventions and policies to improve patient outcomes.

Furthermore, the integration of machine learning into stroke outcome prediction holds great promise for enhancing clinical decision-making and healthcare resource utilization. Developing a machine learning-based prediction tool tailored to the Malaysian population represents an innovative approach to addressing the pressing challenges associated with stroke management. This tool has the potential to transform the way healthcare professionals assess stroke patients, enabling personalized care strategies and optimizing healthcare delivery.

In summary, the public health significance of conducting this study cannot be overstated. Stroke remains a leading cause of disability and mortality worldwide,

imposing a substantial burden on individuals, families, and healthcare systems. In Malaysia, where stroke prevalence is on the rise due to aging population trends and shifts in lifestyle factors, understanding the intricacies of stroke management and prognosis is imperative for mitigating its societal impact. By shedding light on the application of machine learning in stroke research and survival analyses tailored to the Malaysian context, this study has the potential to inform public health policies and interventions aimed at reducing the incidence of stroke, improving patient outcomes, and optimizing resource allocation within the healthcare system. Ultimately, the findings from this research endeavor have the capacity to positively impact population health and well-being, offering hope for a future with reduced stroke-related morbidity and mortality in Malaysia.

1.4 Research question

1. What are the patterns and trends of publications related to the application of machine learning in stroke research?
2. What factors are related to death in acute stroke patients in Malaysia?
3. How well do model-free prediction models predict death in acute stroke patients in Malaysia?
4. What is the design and usability of the web-based stroke mortality prediction tool?

1.5 Objective

1.5.1 General objective

To comprehensively analyse trends in publications related to the application of machine learning in stroke outcome modelling, identify prognostic factors, construct predictive models for mortality among acute stroke patients, and develop a web-based application for stroke mortality prediction using data sourced from multiple stroke centres in Malaysia spanning the years 2016 to 2021.

1.5.2 Specific objective

1. To conduct a bibliometric analysis of the machine learning applications in stroke outcome prediction, with a focus on publications up to the year 2022.
2. To develop a model-based prognostic model for mortality among acute stroke patients admitted in Malaysia between 2016 and 2021.
3. To construct a model-free prediction model for mortality among acute stroke patients admitted in Malaysia between 2016 and 2021.
4. To develop and validate a prototype of a web-based application, the Malaysian Ischemic Stroke Mortality Prediction Tool (MIST), using Python Flask technology, aimed at predicting stroke mortality among acute stroke patients.

1.5.3 Research hypothesis

1. There is a significant change in the publication trends related to machine learning applications in stroke outcome modelling in Malaysia from 2016 to 2022.
2. The selected prognostic factors significantly influence mortality among acute stroke patients in Malaysia from 2016 to 2021.
3. Model-free prediction models provide accurate mortality predictions for acute stroke patients in Malaysia from 2016 to 2021.
4. The web-based application, MIST, demonstrates satisfactory content and face validity, and notably enhances the accuracy of stroke mortality prediction among acute stroke patients in Malaysia.

CHAPTER 2

LITERATURE REVIEW

In this section, the literature related to the study will be appraised and organized according to these subheadings:

- Global stroke burden
- Prognostic and predictive modelling in stroke mortality
- Developing prognostic and predictive models for stroke mortality
- Developing and presenting a web-based prediction tool
- Conceptual framework

2.1 Global stroke burden

Stroke continues to be a significant global health concern. According to the most recent Global Burden of Disease (GBD) 2019 estimates, it ranks as the second leading cause of death and the third leading cause of death and disability combined, measured in terms of disability-adjusted life-years lost (DALYs) (Valery L. Feigin et al., 2021). Shockingly, each year, approximately 6.5 million people succumb to the devastating effects of stroke.

Demographically, stroke impacts a wide range of age groups. Roughly 6.0% of all stroke-related deaths occur in individuals aged 15–49, emphasizing its reach across age brackets. A striking statistic reveals that 34.0% of all stroke-related deaths claim lives under the age of 70. Gender distribution shows that 51.0% of these deaths affect men, while 49.0% affect women.

Ischemic stroke dominates the landscape, constituting 70.0% of all stroke cases. It is characterized by a high risk of long-term recurrence. In the year 2019 alone, ischemic stroke was responsible for a staggering 3.29 million deaths, accounting for 50.3% of stroke-related fatalities and 17.7% of all cardiovascular disease (CVD)-related deaths. The burden extends across age groups, with 2.0% of ischemic stroke-related deaths occurring in individuals aged 15–49 and 19.0% in those under 70. Gender-wise, 48.0% of ischemic stroke-related deaths occur in men, and 52.0% in women.

Intracerebral haemorrhage, another form of stroke, claims the lives of nearly three million individuals annually. Alarmingly, approximately 9.0% of all intracerebral haemorrhage-related deaths affect those aged 15–49, demonstrating its impact across a wide age spectrum. Moreover, 47.0% of these deaths occur in people under the age of 70, further emphasizing the reach of this condition. Gender distribution indicates that 55.0% of intracerebral haemorrhage-related deaths affect men, while 45.0% affect women.

Subarachnoid haemorrhage contributes to the global stroke burden, causing over 373,000 fatalities each year. This condition notably affects younger individuals, with approximately 17.0% of all subarachnoid haemorrhage-related deaths occurring in the 15–49 age group. A substantial 56.0% of these deaths involve individuals under 70. Gender distribution is even, with 50.0% of subarachnoid haemorrhage-related deaths affecting both men and women.

In Malaysia, as in numerous other nations worldwide, stroke, or cerebrovascular disease, stands as the country's third leading cause of mortality. In 2019, Malaysia witnessed 19,928 stroke-related deaths, underscoring the substantial impact of this condition on the nation's health landscape (Valery L. Feigin et al., 2021;

Tan and Venketasubramanian, 2022). While it is noteworthy that Malaysia exhibits lower age and sex-standardized stroke mortality rates and a declining trend in mortality rates for both sexes compared to many other countries in Southeast Asia, a related concern looms large (Hwong et al., 2021; Venketasubramanian et al., 2017). Despite the reduced mortality rates, there exists a significant population of individuals grappling with the debilitating consequences of stroke-related disabilities (Azlin Mohd Nordin et al., 2016).

While Malaysia has made commendable strides in reducing stroke-related mortality rates, the ongoing challenge lies in refining stroke prevention strategies. Recognizing that prevention is often the most effective intervention, there is a growing emphasis on predictive measures to proactively identify individuals at elevated risk of stroke-related complications. In this context, the development and implementation of robust stroke mortality prediction models are pivotal. These models should provide a timely and precise risk assessments. By focusing on mortality prediction, we can not only save lives by enabling targeted interventions but also alleviate the burden of stroke-related disabilities, improving the overall quality of life for stroke survivors.

2.2 Prognostic and predictive modelling in stroke mortality

Prognostic and predictive modelling represent indispensable tools in the healthcare landscape, offering critical insights into disease prognosis and facilitating patient stratification based on their risk profiles. While distinct, these two modelling approaches work collaboratively to inform both patients and healthcare providers about disease progression, support decision-making regarding treatment strategies, optimize healthcare resource allocation, and ultimately contribute to cost-effective healthcare management (Fahey et al., 2018; Vogenberg, 2009; Q. Wang et al., 2022).

The development of prognostic and predictive models inherently entails a multi-faceted process. This process encompasses several crucial steps, including the selection of pertinent candidate predictors, the establishment of suitable data collection methodologies, the meticulous evaluation of data quality (including the handling of missing data), the judicious choice of modelling techniques (ranging from statistical model specifications to machine learning algorithms), and the rigorous assessment of prediction performance. In the development of prediction models, it is imperative to undertake internal validation to gauge model calibration and discrimination, thereby ensuring accurate predictive performance. Furthermore, external validation is essential to assess how well the model generalizes to data not utilized during its initial development phase (Collins et al., 2015; Moons et al., 2009; Royston et al., 2009).

The prediction of stroke mortality has witnessed the emergence of two prominent modelling approaches: model-based and model-free prediction. Both approaches have demonstrated their effectiveness in the realm of stroke primary prevention, leveraging patient characteristics, clinical data, and treatment processes to estimate mortality risk (Collins et al., 2015). Model-based prediction typically employs techniques such as logistic regression and Cox proportional hazard regression, whereas model-free prediction harnesses various machine learning methods, including artificial neural networks (ANN), Naïve Bayes, support vector machines (SVM), decision trees (DT), and random forests (RF) (Fahey et al., 2018; W. Wang et al., 2020).

Model-based prediction methods often rely on a priori statistical assumptions, such as variable independence and specific probability distribution assumptions, which can provide a structured framework for analysis. These methods typically require the outcome variable to be binomial and may offer interpretability of model parameters. However, they may oversimplify complex relationships in the data and

may not capture nonlinear associations effectively. Conversely, model-free approaches adapt to inherent data properties without preconceived modeling assumptions, enabling the creation of non-parametric representations through machine learning algorithms. This flexibility allows for capturing intricate patterns and interactions within the data, leading to potentially higher predictive accuracy. However, model-free methods may be computationally intensive, require large datasets, and lack interpretability compared to model-based approaches (Gao et al., 2018).

Both model-based and model-free prediction methods offer unique advantages and disadvantages in the exploration of stroke mortality. Combining the statistical rigor of model-based approaches with the flexibility of machine learning-based model-free methods allows for a comprehensive examination of stroke outcomes from different angles. By leveraging the strengths of each approach, researchers can gain a deeper understanding of the complex factors influencing stroke mortality and develop more robust prediction models that improve patient care and clinical decision-making.

Ensuring transparency and rigorous reporting in the development and validation of prediction models remains paramount. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline serves as an indispensable resource in this endeavour and finds widespread application across various medical domains (Collins et al., 2015). Originally tailored for regression modelling, TRIPOD has since evolved to accommodate the development, validation, and updates of models employing machine learning methods. An adjusted TRIPOD statement, presented in Appendix D, ensures that studies involving prognostic modelling adhere to stringent reporting standards. This checklist serves as an invaluable guide to communicate the findings of this study accurately and transparently.

2.3 Developing prognostic and predictive models for stroke mortality

The initial phases of model development encompass several crucial steps. These include encoding the predictors, defining the model's specifications, and executing the estimation process. Subsequently, we delve into evaluating the prediction model's performance, considering facets such as discrimination and calibration. Following this assessment, we explore methodologies for both internal and external validation, as well as strategies for model refinement and updating. Lastly, we address the pivotal aspects of model reporting and presentation, concluding with a discussion of the steps involved in the practical implementation of the model within clinical settings.

In summary, the model development journey entails a comprehensive process that encompasses predictor encoding, model specification, estimation, performance evaluation, validation, refinement, reporting, presentation, and practical implementation. These sequential steps are fundamental to ensuring the robustness and clinical utility of predictive models in healthcare and medical research.

2.3.1 Predictors of stroke mortality

The prediction of stroke mortality relies on amalgamating multiple predictors and assigning relative weights to each of these predictors to calculate the risk or probability of stroke-related death. Among the routinely collected factors, several variables have consistently emerged as key predictors of functional outcomes and mortality in stroke patients. These variables encompass age, sex, disease characteristics such as severity and subtype, intravenous intervention, and comorbid conditions including diabetes mellitus and atrial fibrillation (Fahey et al., 2018).

Furthermore, it is crucial to consider various other medical conditions that significantly contribute to an elevated risk of stroke mortality. One such condition is chronic kidney disease, which has been shown to cause a three-fold increase in

mortality rates among stroke patients (Bobot et al., 2023). Additionally, a history of hypertension plays a pivotal role in drastically elevating the risk of stroke mortality, with hazard ratios ranging from as low as 1.47 to as high as 9.27 times greater risk (G. Hu et al., 2005). Moreover, the presence of heart disease, particularly heart failure, is associated with nearly two to three times higher stroke mortality risk (Pana et al., 2019). Smoking, a well-known contributor to cardiovascular and cerebrovascular diseases, also leads to an increased risk of stroke mortality (Hou et al., 2017). However, it's noteworthy that the presence of high cholesterol levels is directly related to an increased risk of stroke mortality (Olsen et al., 2007).

While these factors are less frequently incorporated into most stroke mortality prediction models or stroke-related registries, they hold significant importance in understanding and mitigating the risk of stroke mortality. Factors such as poor premorbid conditions, including undernutrition and a low premorbid Modified Rankin Scale (mRS), have been found to be associated with a higher risk of stroke mortality (Davis et al., 2004; Han et al., 2020). Furthermore, it's not only patient-related factors that should be considered; caregiver-related factors also play a role in increasing the risk of stroke mortality. For instance, caregiver stroke education and mental health can significantly contribute to this risk, with anxiety and depression among main family caregivers being associated with a higher risk of 6-month mortality among patients with moderate to severe strokes (Hong et al., 2017; Zhao et al., 2021). Understanding the multifaceted nature of these risk factors is essential for comprehensive stroke mortality prediction and improved patient care.

Various prognostic models have been developed to predict stroke mortality, each leveraging a unique combination of these influential predictors. Notable models in this domain include the Ischemic Stroke Predictive Risk Score (ISCORE), Totaled

Health Risks in Vascular Events (THRIVE) score, the Get With The Guideline (GWTG) based score, Essen Stroke Risk Score (ESRS), Stroke Subtype, Oxford Community Stroke Project Classification, Age, Pre-stroke Modified Rankin (SOAR) score, and the Preadmission comorbidities, Level of consciousness, Age, and Neurologic deficit (PLAN) score (Flint et al., 2015; Gent, 1996; Myint et al., 2014; O'Donnell et al., 2012; Saposnik et al., 2011; Smith et al., 2010; Christian Weimar et al., 2009). These models have been instrumental in identifying and quantifying the importance of various predictors for stroke mortality outcomes.

To provide a comprehensive overview of the existing models and their methodologies, **Table 2.1** has been included, summarizing pertinent details such as data sources, predictor variables utilized, and the analytical approaches employed in previous stroke mortality prediction scores.

Table 2.1 Summary of data source, predictors, and analysis used in previous stroke mortality prediction models

Prediction model	Data Source	Predictors	Analysis
ISCORE (Saposnik et al., 2011)	Registry of the Canadian Stroke Network and Ontario Stroke Audit	Age, sex, Risk factors, stroke severity, stroke subtypes, comorbid condition, preadmission disability and glucose on admission	Multiple logistic regression
THRIVE score (Flint et al., 2015)	Mechanical Embolus Removal in Cerebral Ischemia (MERCI) trial	National Institutes of Health Stroke Scale (NIHSS), age, and presence of hypertension (HTN), diabetes mellitus (DM), or atrial fibrillation (AF)	Multiple logistic regression
GWTG based score (Smith et al., 2010)	Internet-based hospital registry	Age > 60 years, atrial fibrillation, coronary artery disease, diabetes mellitus, and peripheral vascular disease	Multiple logistic regression
ESRS (Gent, 1996; Christian Weimar et al., 2009)	19 185 patients from 384 clinical centres in a randomised, blinded, trial of Clopidogrel versus Aspirin in Patients at Risk of Ischaemic Events (CAPRIE)	Age 65-75 years, age >75 years, Arterial hypertension, diabetes mellitus, previous myocardial infarction (MI), other cardiovascular disease (except MI and atrial fibrillation), peripheral arterial disease, smoker, previous transient ischaemic attack (TIA)	Score performances were evaluated with area under the curve (AUC) by c-statistic and calibration $\chi 2$ (survival modified Hosmer-Lemeshow)
SOAR (Myint et al., 2014)	United Kingdom stroke registry	Age category, gender, stroke type, Oxfordshire Community Stroke Project (OCSP) classifications, pre-stroke modified Rankin scale.	Multiple logistic regression
PLAN (O'Donnell et al., 2012)	Registry of the Canadian Stroke Network	Preadmission comorbidities, level of consciousness, age, neurological deficit.	Multiple logistic regression

2.3.2 Issues with model development

2.3.2(a) Addressing missing data through imputation techniques

Developing prediction models is a crucial endeavour in various fields, enabling us to make informed decisions based on available data. However, missing data can pose significant challenges to the development of accurate and reliable prediction models. To address these challenges, it is essential to understand the mechanisms and implications of missing data in the context of prediction model development.

2.3.2(a)(i) Type and effect of missing data

The influence of missing data on the process of developing prediction models hinges on the underlying mechanisms responsible for data absence, with three primary mechanisms being identified: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Dziura et al., 2013; Little et al., 2012; Sterne et al., 2009).

In the case of MCAR, data is missing entirely at random, and the resulting incomplete datasets maintain their representativeness for the entire dataset. This particular form of missingness primarily exerts its influence on the model's statistical power. It is important to note that MCAR entails data absence independent of both observed and unobserved data. While MCAR itself does not introduce bias, it does have the potential to diminish the model's statistical power due to the resultant reduction in sample size (Little et al., 2012).

Conversely, in the context of MAR, missingness is contingent upon observed data but not influenced by the nature of the missing data itself. MAR provides an opportunity to predict missing values based on the information available from participants with complete data, thereby safeguarding the overall integrity of the model.

This predictive approach allows for the preservation of the utility of the data for model development (Dziura et al., 2013).

In the case of MNAR, missing data is not randomly distributed but rather depends on the nature of the missing data itself, even when considering observed data. MNAR presents a substantial challenge to the development of prediction models due to its potential introduction of bias. Addressing MNAR necessitates a meticulous examination and the implementation of sensitivity analyses to evaluate its potential impact on model performance (Sterne et al., 2009).

To summarize, the type and effect of missing data are critical considerations in prediction model development, as they influence both the statistical power of the model and the integrity of its predictions. Understanding the underlying mechanisms behind missing data—MCAR, MAR, or MNAR—provides essential insights for researchers to navigate the complexities of prediction modelling effectively.

2.3.2(a)(ii) Handling missing data

In prediction model development, single or multiple imputation is a widely accepted approach for handling missing data (Zhang et al., 2017). This technique involves imputing single or multiple sets of plausible values for missing data based on an imputation model (S van Buuren et al., 1999; Stef van Buuren and Groothuis-Oudshoorn, 2011). These imputed datasets are then used for model development and evaluation.

Sensitivity analyses are also crucial in prediction model development to assess the robustness of the models under different assumptions about missing data (Little et al., 2012; Morris et al., 2014). These analyses help researchers understand how variations in handling missing data may affect the predictive performance of the model.

In conclusion, addressing missing data in prediction model development is essential for ensuring the accuracy and reliability of the models. Careful consideration of the missing data mechanisms, appropriate imputation techniques, and sensitivity analyses are fundamental steps in mitigating the impact of missing data on the development of predictive models.

2.3.3 Modelling approaches

2.3.3(a) Model-based prediction of acute stroke mortality: time-to-event modelling

Epidemiological studies, especially cohort study designs, involve a meticulous and extended observation of participants over a predefined timeframe. In this setup, participants are closely monitored, and during the course of the study, they may either experience the event of interest (referred to as "failure") or continue to be under observation without experiencing the event (referred to as "censoring"). Additionally, in cohort studies, there are situations where participants may have consistent follow-up durations, while in other instances, their follow-up times are individualized, each having its specific start and end points. In cases where individualized follow-up times are prevalent, survival analysis emerges as the most suitable and preferred data analysis strategy (Kleinbaum and Klein, 2012).

To delve deeper, a cohort study typically assembles a group of individuals who share a common characteristic or exposure, and then they are monitored over time to assess specific health outcomes. Throughout this observation period, some participants may encounter the outcome or event under investigation (termed "failure"), while others may remain unaffected by it but continue to be observed (referred to as "censoring"). This differentiation between participants who experience the event and those who do not is pivotal for conducting survival analysis.

Significantly, in cohort studies, the concept of uniform or individualized follow-up times becomes relevant. In situations where all participants undergo the same follow-up duration, the analysis is simplified as the entire cohort is observed for a fixed period. However, in many real-world scenarios, individual participants exhibit diverse follow-up times due to variations in entry times, event occurrences, or other factors. In such cases, the application of survival analysis is imperative.

Survival analysis, in this context, enables researchers to accommodate these varying follow-up times and event occurrences over time. It accommodates the dynamic nature of cohort studies, where participants may enter or exit the study at different junctures and experience events at different times. By employing survival analysis, researchers can effectively analyse data from cohort studies with individualized follow-up times, providing a robust and comprehensive approach to comprehend and interpret epidemiological research outcomes.

In the realm of traditional survival analysis, a fundamental statistical technique in medical and epidemiological research, three primary approaches prevail: non-parametric, semi-parametric, and parametric survival analysis. These distinct model-based prediction methods offer a range of tools for comprehending and analysing survival data (Kleinbaum and Klein, 2012; Schwender, 2012).

Within the domain of non-parametric survival analysis, the Kaplan-Meier method assumes significance as a valuable tool. This method excels in depicting survival probability over time through the creation of survival curves. These curves offer a visual representation of how survival probabilities change across distinct time intervals, making them a potent descriptive tool for understanding survival patterns.

Semi-parametric techniques, like the Cox proportional hazard regression, have become a cornerstone in the analysis of survival data in epidemiology. This approach provides a flexible framework for evaluating the influence of covariates on survival outcomes while making minimal assumptions about the underlying hazard function. The Cox proportional hazard regression is widely utilized to gauge the hazard ratios associated with various factors, yielding valuable insights into how different variables affect survival times.

In parametric survival analysis, models such as the Weibull, Exponential, and lognormal survival models come into play. Although less commonly employed than non-parametric and semi-parametric approaches, parametric models offer distinct advantages. However, they require specifying the baseline hazard function, rendering them more reliant on specific assumptions. These models can be particularly beneficial when the underlying distribution of survival times aligns with one of these parametric forms, allowing for precise estimation of survival probabilities.

A comprehensive systematic review and meta-analysis of clinical prediction models for stroke, encompassing 35 years of literature and analysing 109 articles, unveiled that survival analysis, particularly the Cox proportional hazard regression, ranks as the second most frequently employed analysis method for predicting stroke outcomes, while parametric survival analysis is less commonly utilized (Fahey et al., 2018). **Table 2.2** summarizes the key findings from this study.

Table 2.2 Summary of findings from systematic review and meta-analysis of stroke clinical prediction models (Fahey et al., 2018)

Characteristics	Number of models
Total	66 models
Outcome	
Mortality	27 models
Function	28 models
Mortality/function	11 models
Source of data	
Randomized control trial	36 models
Registry data	31 models
Model development	
Logistic regression	62 models
Cox regression	10 models
General estimating equations	4 models
Linear models	2 models
Data mining	2 models
Variables selection methods	
Backward	25 models
Forwards	20 models
Not specified	21 models
Internal validation	
Random split	30 models
Bootstrap	39 models
Cross validation	1 model
External validation	
Temporal	22 models
Geographical	40 models
Different setting	2 models
Different investigators	0 model

2.3.3(b) Model-free Prediction of Acute Stroke Mortality: survival machine learning modelling

As cutting-edge computing technology continues to advance, coupled with the availability of model-free prediction models, machine learning methods have emerged as highly promising tools for handling high-dimensional data. These methods exhibit remarkable capabilities in identifying nonlinear patterns and optimizing outcome predictions. In the realm of stroke research, machine learning techniques are increasingly being harnessed to predict both short-term and long-term stroke

outcomes, encompassing aspects such as functional status, mortality, or a combination of these factors (Alaka et al., 2020; Heo et al., 2019; Scrutinio et al., 2020).

Notably, a repertoire of five prominent learning algorithms, including DeepSurv, Cox-Time, Cox model with Elastic Net (Cox-EN), Support Vector Machine (SVM), and Random Survival Forest (RSF), have gained traction for time-to-event machine learning modelling (Ishwaran et al., 2008; Katzman et al., 2018; Kvamme et al., 2019; Van Belle et al., 2007; Zou and Hastie, 2005).

DeepSurv, a remarkable innovation in this landscape, represents a Cox proportional hazards (CPH) feed-forward deep neural network model. This model is designed to generate non-linear risk representations for clinical events based on input features (Katzman et al., 2018). Its architecture comprises a patient data input layer, a hidden layer with fully connected nodes and a dropout layer, culminating in a single node with linear activation that estimates the log-risk function within the CPH model. What sets DeepSurv apart is its ability to provide predictions without necessitating the specification of interaction terms, a feat unattainable with traditional models. Furthermore, the flexibility of DeepSurv allows for dynamic adjustments of model hyperparameters based on performance.

Cox-Time introduces a significant departure from traditional Cox models by offering a relative risk algorithm based on neural networks. This innovation liberates the model from the proportionality assumption that constrains the traditional approach. Operating in continuous time, much like the DeepSurv algorithm, Cox-Time introduces a time-dependent covariate to accommodate non-proportional variables, thus enriching the modelling toolbox (Kvamme et al., 2019).

Certainly, Elastic Net (EN) can be seamlessly integrated into the Cox proportional hazards model to serve as a potent feature selection technique (Xiao et al., 2022). The Cox proportional hazards model, a cornerstone of survival analysis, is adept at assessing the impact of covariates on the hazard function. However, in scenarios where datasets are characterized by high dimensionality or a multitude of potential predictor variables, the imperative to discern the most pertinent and informative features becomes apparent.

EN offers a compelling solution by harmonizing the L1 and L2 penalties inherited from the Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression techniques (Zou and Hastie, 2005). This synthesis achieves a delicate equilibrium between LASSO's capacity to induce sparsity and ridge regression's prowess in regularization. Such synergy empowers EN to navigate the intricacies of multicollinearity among predictor variables while concurrently performing astute feature selection.

Incorporating EN into the Cox model confers a spectrum of advantages. Firstly, it undertakes the task of feature selection with finesse, endowing non-zero coefficients to pertinent features while relegating less influential ones to a coefficient of zero. This judicious feature selection process streamlines the model, rendering it more interpretable. Furthermore, the regularization effect engendered by EN safeguards the Cox model against overfitting, a pertinent concern when dealing with high-dimensional data. By mitigating the proclivity to overfit, the model becomes more adept at generalizing to new data, thus bolstering its predictive performance. Lastly, the Cox-EN amalgamation enriches the model's interpretability, as it focuses attention on a subset of salient features, thereby facilitating a deeper understanding of the factors that underpin survival outcomes (Zou and Hastie, 2005).