UNIVERSITI TEKNOLOGI MARA

JIIICaS 2024
JOHOR INTERNATIONAL INNOVATION, INVENTION COMPETITION AND SYMPOSIUM

PROCEEDINGS OF JOHOR INTERNATIONAL INNOVATION INVENTION COMPETITION AND SYMPOSIUM 2024 (JIIICaS 2024)

"Flourish and Nurturing Sustainable Innovation for a Prosperous Nation"

JIICOS 2024
JOHOR INTERNATIONAL INNOVATION, INVENTION COMPETITION AND SYMPOSIUM

Editorial Board

e ISBN: 978-967-0033-25-9

Published in Malaysia by
Universiti Teknologi MARA Cawangan Johor
Kampus Pasir Gudang
81750 Masai

Preface

In the name of Allah, the Almighty who gives us the enlightenment, the truth, the knowledge and with regards to Prophet Muhammad (peace be upon him) for guiding us to the straight path. We thank to Allah for giving us guidance and strength to write this e-book.

This e-book compiles the extended abstracts that submitted to Johor International Innovation Invention Competition and Symposium 2024 (JIIICaS2024), where JIIICaS2024 is a virtual platform for all creative minds to share and present their invention and innovation. Each abstract gives a brief background on the innovation or project.

We hope that this e-book will help the readers to get to know the innovation done by the students and get some ideas to develop future innovation products.

Foreword Rector

Assalamualaikum warahmatullahi Wabarakatuh, Salam Sejahtera, Salam Malaysia MADANI and Salam UiTM Dihatiku.

In the name of Allah, the Most Gracious, the Most Merciful.

It is a great honor to welcome you to the Johor International Innovation, Invention, Competition, and Symposium 2024 (JIIICaS 2024). This event connects various disciplines, focusing on education and engaging educators, students, researchers, and innovators from all walks of life.

Innovation is not just about ideas; it demands perseverance, creativity, and determination to turn those ideas into reality. The remarkable projects showcased today highlight the dedication and spirit of all participants. Initiatives like this not only explore new technologies but also cultivate skills and leadership among our youth. At Universiti Teknologi MARA (UiTM) Johor Branch, we are fully committed to fostering a dynamic culture of innovation, promoting the commercialization of new products, and encouraging meaningful collaborations with industry and society.

As we celebrate this event, I would like to extend my heartfelt gratitude to all sponsors, judges, the College of Computing, Informatics and Mathematics, UiTM Pasir Gudang Campus as the event organizer, as well as to the researchers and participants for their hard work in making this event a success. Let us continue striving for innovation and excellence. May the ideas presented today inspire us and lay the groundwork for future achievements.

Thank you.

Associate Professor Dr. Saunah Zainon
Rector
Universiti Teknologi MARA (UiTM)
Johor Branch

(A-ST176) PREDICTION OF BRAIN STROKE OCCURRENCE USING SUPPORT VECTOR MACHINE

Amzar Raziq Khan Bin Azam Khan
Bachelor of Computer Science (Bioinformatics)
Universtiti Teknologi Malaysia-
Johor Bahru, Malaysia
amzar2000@graduate.utm.my

Dr. Izyan Izzati Binti Kamsani
Lecturer / Supervisor
Universiti Teknologi Malaysia-
Johor Bahru, Malaysia
izyanizzati@utm.my

ABSTRACT

Stroke is a serious medical condition identified by an abrupt interruption of blood flow to the brain. Risk factors for stroke include high blood pressure, diabetes, and smoking. Due to their reliance on static risk factors and limited predictive abilities, traditional methods of predicting strokes may not be as effective today because they frequently ignore dynamic factors and individual variations in disease progression and response to treatment. Machine learning has demonstrated the potential for identifying unknown threats, such as brain stroke. This research project focuses on using supervised machine learning methods to improve stroke prediction's precision and efficiency. The chosen method is Support Vector Machine (SVM), which is a popular supervised machine learning method. This research aims to create a predictive model for brain strokes by comparing performance measurement of supervised machine learning Support vector machines (SVM) using a brain stroke dataset. This research used a structured methodology to develop a supervised learning model for predicting brain strokes. The process has 3 phases which involves literature review, data gathering, pre-processing, model development using SVM, and performance evaluation. The performance on the known stroke dataset is then analysed, and the ability to identify brain stroke is evaluated. The results show that the selected classifier effectively detect and predict brain stroke and outperform existing techniques. The results of this research will help in the development of reliable predictive models to evaluate the risk of stroke, allowing healthcare professionals to quickly intervene and put preventive measures into place. Ultimately, the integration of supervised machine learning techniques has the potential to increase the accuracy of stroke prediction and enable early intervention, thereby easing the burden of stroke-related morbidity and mortality.

Keywords: brain stroke, supervised machine learning algorithm, support vector machine

I. Introduction

Stroke is a leading cause of disability and death worldwide. The World Health Organization states that stroke is the second leading cause of death globally, responsible for 11% of total deaths [1]. Stroke occurs when blood flow to the brain is disrupted, causing brain cells to die from a lack of oxygen and nutrients, leading to symptoms like weakness, numbness, vision issues, speech difficulty, and cognitive impairment.

Early identification and treatment of stroke are crucial for improving patient outcomes, but identifying high-risk individuals remains challenging. Traditional risk scoring systems like the Framingham Risk Score have limited accuracy and may not account for the complexity and variation of risk factors among individuals. Machine learning (ML) offers significant promise in predicting stroke risk and improving diagnosis and treatment. ML algorithms can analyze large datasets to uncover patterns and relationships not easily seen by humans. Studies have shown that supervised ML algorithms can effectively predict stroke risk factors and outcomes. For example, Attia [2] found that ML algorithms outperformed traditional risk scoring systems in predicting stroke in individuals with atrial fibrillation.

Brain stroke can be classified into two main types: ischemic and hemorrhagic. Ischemic stroke, accounting for about 87% of all cases, occurs when a blood clot blocks a blood vessel in the brain, leading to symptoms like weakness, numbness, and difficulty speaking. Hemorrhagic stroke, making up approximately 13% of stroke cases, happens when a blood vessel in the brain bursts, causing symptoms such as a severe headache, nausea, and loss of consciousness [3]. Strokes can be fatal or cause significant disability, with symptoms including sudden numbness or weakness, confusion, difficulty speaking, vision problems, dizziness, and loss of coordination. Early recognition and treatment are crucial for improving outcomes. With an estimated 15 million cases worldwide annually, stroke is a major public health challenge [4]. Identifying high-risk individuals is difficult due to numerous risk factors. Machine learning (ML) algorithms have shown promise in predicting medical conditions, including stroke. This research project aims to develop a predictive model for brain stroke using supervised ML, which learns from labeled training data to identify patterns and predict outcomes. The model will be trained on a dataset containing various risk factors, such as age, sex, hypertension, and heart disease. Once trained, the model will predict stroke likelihood in new individuals and be tested for accuracy. This approach could enhance stroke prevention and early intervention, leading to better patient outcomes and reduced healthcare costs. Additionally, it may aid in developing more accurate prediction models for other medical conditions.

II. Literature Review

Brain stroke, or cerebrovascular accident (CVA), occurs when blood supply to the brain is interrupted, leading to tissue damage. It can be ischemic (caused by a blood clot) or hemorrhagic (caused by a burst blood vessel). Stroke is a leading cause of disability and death worldwide, with early diagnosis and treatment crucial for better outcomes.

Hypertension, smoking, and diabetes are major risk factors for stroke, accounting for over 60% of cases in China [5]. In 2016, there were 116.4 million stroke cases globally, with 42.2 million survivors living with disabilities [6]. Diagnosis involves physical exams, blood tests, CT or MRI scans, and other tests to assess brain damage. Treatments include surgery, medication, and rehabilitation. Early recognition and medical attention are essential for improving outcomes.

In medical research, machine learning (ML) has become an effective tool due to its better data processing and prediction skills over conventional techniques. Through the analysis of patterns regarding risk factors like age, gender, hypertension, and heart disease, supervised machine learning (ML) which learns from labeled data can predict the probability of stroke. High accuracy in predicting medical outcomes, such as stroke, has been demonstrated using ML models. The goal of this project is to create a machine learning (ML) predictive model for stroke that could improve patient outcomes overall, early intervention, and prevention. Technological developments of this kind may potentially enhance prediction models for other illnesses.

Machine learning (ML) plays a crucial role in diagnosing, treating, and managing diseases by analyzing vast data, including genetic information, medical imaging, and electronic health records, to find patterns and make predictions. The ability of ML to improve disease diagnosis is significant, with algorithms designed to predict the likelihood of diseases such as diabetes, cancer, and heart disease using patient data like age, sex, and medical history. For example, a study published in Diabetes Care demonstrated an ML model accurately predicting a patient's risk of developing diabetes over the next five years [7].

A. Supervised Machine Learning

Supervised machine learning is a type of machine learning where an algorithm is trained on a labeled dataset, meaning each training example includes input data and the corresponding correct output. The algorithm may transform inputs into outputs by looking for patterns in the data. Support vector machines, xgboost, and random forests are popular supervised learning algorithms for tasks like regression and classification. Following training, the model predicts the outcome for new, unidentified data using the understanding it has acquired from the labeled samples.

- Support Vector Machine (SVM) is a supervised learning algorithm used for classification by finding the optimal hyperplane that separates different classes in a high-dimensional space.

- XGBoost is an advanced gradient boosting algorithm that builds an ensemble of decision trees, improving performance through techniques like regularization and parallel processing.

- Random Forest is an ensemble method that constructs multiple decision trees from random data subsets, averaging their predictions to enhance accuracy and reduce overfitting.

Support Vector Machine (SVM) is a powerful and effective machine learning algorithm used primarily for classification tasks, though it can also be adapted for regression. SVM finds the optimal hyperplane that maximally separates different classes in a high-dimensional feature space. It uses kernel functions to handle non-linear decision boundaries, making it versatile for various applications. In medical diagnostics, SVM has shown promising results, such as in predicting strokes with high accuracy, outperforming other models like Logistic Regression and K-Nearest Neighbors [8].

SVM's ability to handle high-dimensional data and its resistance to overfitting make it a robust choice for complex prediction problems. By utilizing different kernel functions, SVM can adapt to various data characteristics, providing flexibility and precision in model building [9].

III. METHODOLOGY

A. Research Methodology

This sector outlines the research framework, summarizing essential tasks for maintaining a consistent research schedule. The framework comprises three phases: literature review and problem determination, stroke dataset gathering and preprocessing, and supervised learning model development, focusing on SVM classifier. The final phase involves testing, evaluation, and comparison of the classification method's effectiveness with previous studies' methods. Figure 1 illustrates the research methodology's phases for clarity.

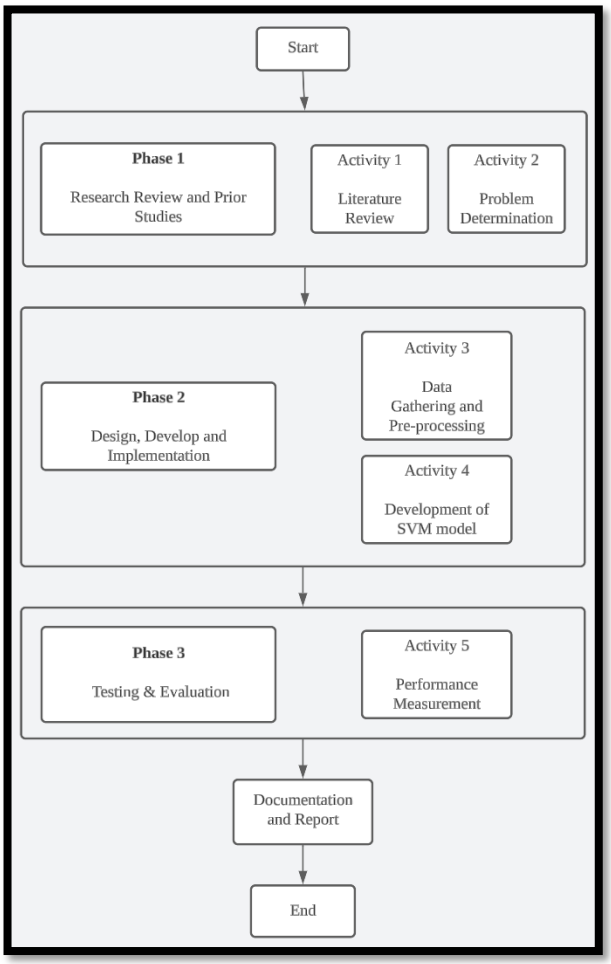


Figure 1 Research Framework

1) Phase 1: Research Review and Prior Studies

In this phase, a great deal of work was done to gather and review pertinent studies and data from a variety of sources, including trustworthy books, journals, and articles from websites like as ResearchGate, NCBI, and IEEE. To define the research's aim and comprehend the benefits and drawbacks of previous studies' approaches, a comprehensive review was undertaken. This procedure includes assessing domain-specific expertise, tools, algorithms, and previous research to find different machine learning approaches and techniques for stroke diagnosis. This exhaustive analysis ensured a clear knowledge of the topic at hand and helped uncover gaps or concerns that needed to be addressed, providing a solid foundation for the following research phases.

2) Phase 2: Design, Develop and Implementation

This phase involves gathering the dataset, preprocessing the data, selecting features, and developing the models. The brain stroke data in the dataset, which was gathered from Kaggle, covers a variety of characteristics and samples. To guarantee reliability and accuracy, this data is processed and any missing values are handled.

The model's top features are determined and chosen. A number of machine learning models are created and refined for predicting brain stroke, such as Support Vector Machine (SVM).

3) Phase 3: Testing & Evaluation

In this phase, the developed machine learning models are evaluated using a separate dataset to ensure unbiased performance assessment. Cross-validation is used to validate the models during the evaluation phase, and metrics including accuracy, precision, recall, and F1 score are calculated. To show progress and identify areas that want improvement, a comparison with baseline models and prior research is done. To achieve better results, the models are tuned and adjusted for improved performance. Ensuring the models are reliable, accurate, and prepared for actual use in the prediction of brain stroke is the goal.

B. Performance Measurement

The accuracy, F1-score, precision, and recall of the performance measurement for stroke prediction are used in this study. Other than that, confusion matrix will be used to accurately measure the performance of the machine learning models. Particularly the SVM models.

1) Confusion Matrix

A confusion matrix is a performance evaluation tool used in machine learning to assess the accuracy of a classification model. It displays a tabular comparison between the model's predicted values and the dataset's actual values. The actual class labels are represented by each row in the matrix, while the expected class labels are represented by each column. The matrix's diagonal elements show how many predictions were accurate for each class, while the off-diagonal elements show how many misclassifications there were. This makes it possible to analyze the model's performance in-depth using measures like accuracy, precision, recall, and F1-score, which helps to identify the model's advantages and disadvantages for class prediction. The confusion matrix's tabular representation is displayed in the figure below.

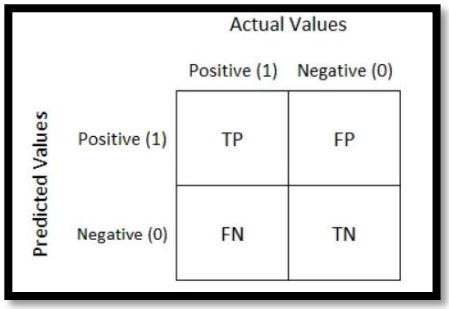


Figure 2 Confusion Matrix

- True Positive (TP): The number of instances where the model correctly predicted the positive class.
- False Positive (FP): The number of instances where the model incorrectly predicted the positive class (also known as Type I error).
- False Negative (FN): The number of instances where the model incorrectly predicted the negative class (also known as Type II error).
- True Negative (TN): The number of instances where the model correctly predicted the negative class.

2) Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a method used to address class imbalance in machine learning datasets by generating synthetic examples for the minority class. It works by selecting a minority class sample, finding its k-nearest neighbors, and creating new data points by interpolating between the sample and its neighbors. This process balances the class distribution, improving model performance by providing a more representative sample and reducing overfitting compared to simple duplication. In applications like stroke prediction, SMOTE helps model better identify and predict occurrences by ensuring the minority class is adequately represented in the training data.

IV. EXPERIMENT

A. Proposed Model

To ensure systematic research execution, the study followed three phases outlined in Chapter 3. Phase 2's development and experimental frameworks commenced in this chapter, starting with acquiring the patient stroke dataset from the Kaggle repository. Initial data processing involved handling missing values and normalizing the dataset to facilitate desired analysis outcomes. Feature selection was conducted as a crucial step in machine learning development, filtering dataset attributes to choose the most relevant features for the model. The Support Vector Machine (SVM) classifier, selected for prediction, received these optimized attributes as inputs.

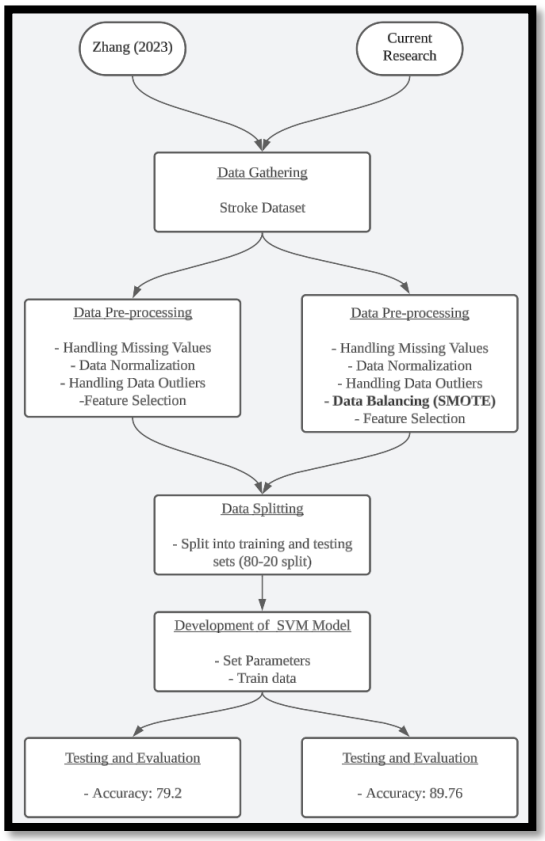


Figure 3 Zhang's and current research flowchart comparison

Figure 3 compares the methodology used in Zhang's (2023) study with the current research on stroke prediction using SVM. While both studies start with data gathering from a stroke dataset, Zhang's study involved basic pre-processing, normalization, and feature selection, resulting in 79.2% accuracy. In contrast, the current research includes additional steps such as handling outliers, data balancing using SMOTE, and feature selection, leading to a higher accuracy of 89.76%. This comparison highlights the effectiveness of the advanced methodologies used in the current research.

B. Data Preparation

Data preparation is crucial for the suggested supervised learning prediction. The original dataset comprises 5110 entries with 12 attributes, indicating a stroke occurrence rate of 5%. However, the data distribution is heavily skewed, with a null accuracy score of 95%, implying that a random prediction model could achieve similar accuracy. To optimize results, oversampling or undersampling during modeling and training data is necessary.

C. Data Pre-processing

Data pre-processing refers to the steps and techniques involved in preparing raw data for analysis or machine learning tasks. It seeks to transform the data into a

suitable format that algorithms and models may use efficiently and accurately. Data inconsistencies, missing values, outliers, and other issues with data quality can negatively affect the performance of analytical or predictive models. Data pre-processing is a vital step in addressing these problems.

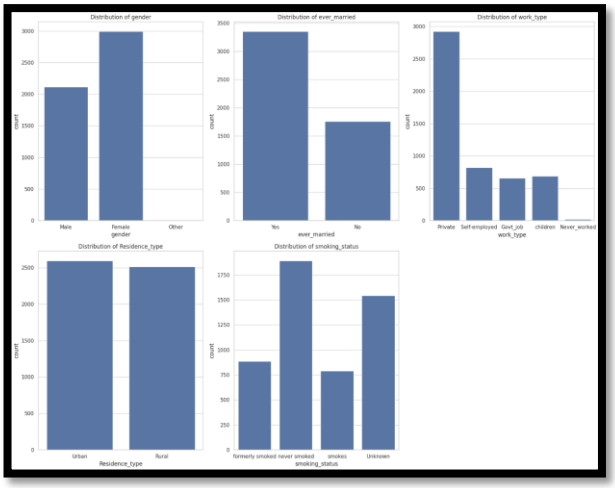


Figure 4 Categorical Data Counts

1) Handling Missing Values

Missing values were found in the raw data gathered for this research, especially in the 'bmi' property, as Figure 5 shows. This problem may cause model predictions to be inaccurate, requiring a data validation procedure. Common methods for dealing with missing values include deleting rows that include missing data, filling in the missing values with the column mean, or leaving them empty or with a specified meaning, such as "Unknown." The suggested solution for this study is to use the column mean in place of any missing "bmi" values in order to guarantee the reliability and precision of the analysis.

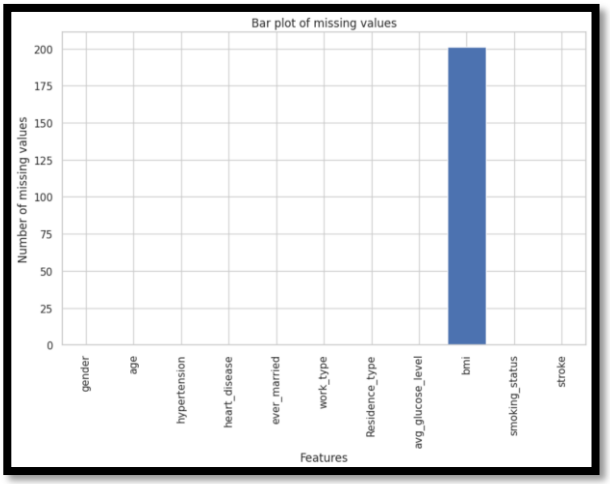
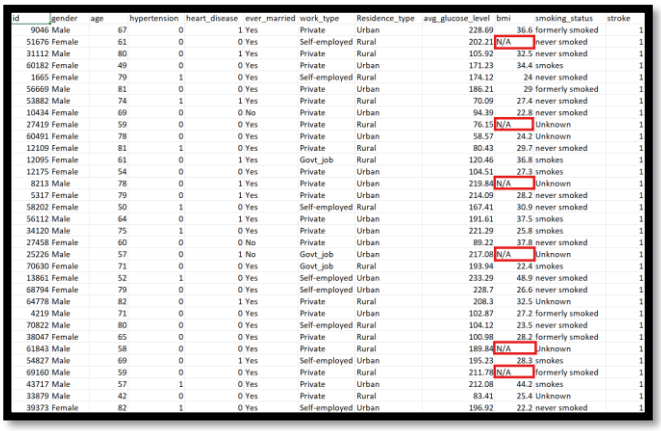


Figure 4 Missing Values



id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1
70822	Male	80	0	0	Yes	Self-employed	Rural	104.12	23.5	never smoked	1
38047	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly smoked	1
61843	Male	58	0	0	Yes	Private	Rural	189.84	N/A	Unknown	1
54827	Male	69	0	1	Yes	Self-employed	Urban	195.23	28.3	smokes	1
69160	Male	59	0	0	Yes	Private	Rural	211.78	N/A	formerly smoked	1
43717	Male	57	1	0	Yes	Private	Urban	212.08	44.2	smokes	1
33879	Male	42	0	0	Yes	Private	Rural	83.41	25.4	Unknown	1
39373	Female	82	1	0	Yes	Self-employed	Urban	196.92	22.2	never smoked	1

Figure 6 Missing Values in Dataset

Figure 6 shows the spreadsheet of the Dataset, the highlighted cells are the missing values which consists of 201 rows. After replacing the missing values with mean values, we have handled any missing values in the dataset as shown in Figure 6.

```
gender              0
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
```

Figure 7 Missing Values after replacing with mean value

2) Data Outliers

Identifying and addressing data outliers is crucial to prevent skewed results and inaccuracies in statistical analyses and machine learning models. In this section, outliers in the 'avg_glucose_level' and 'bmi' columns were detected using the interquartile range (IQR) method, defining outliers as data points falling outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR. For 'avg_glucose_level', 627 outliers were identified with limits set at 21.9775 and 169.3575, while 'bmi' had 126 outliers with limits of 10.3 and 46.3. After adjusting rows containing outlier values, the dataset was free from outliers, ensuring accurate analysis and model performance.

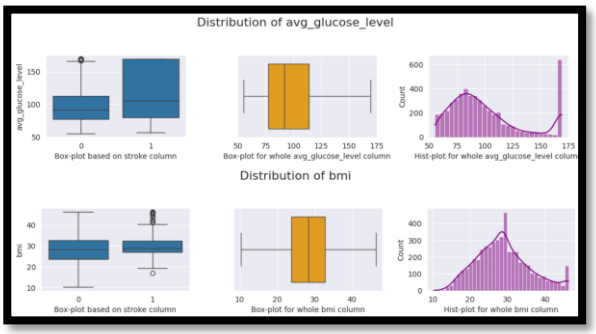


Figure 8 Distribution of avg_glucose_level and bmi after handling data outliers

3) OneHotEncoding

One-Hot Encoding is a technique used to convert categorical data into a numerical format for machine learning algorithms. Each categorical value is converted into a binary vector where a '1' denotes the category's existence and '0's elsewhere; this ensures that no determined ordinal relationship exists between categories. For algorithms that need numerical input to analyze and use category data efficiently, this technique is crucial. One-hot encoding of categorical information in stroke prediction models, such as gender or type of habitation, guarantees that the model can use this data appropriately, increasing its predictive power and overall accuracy.

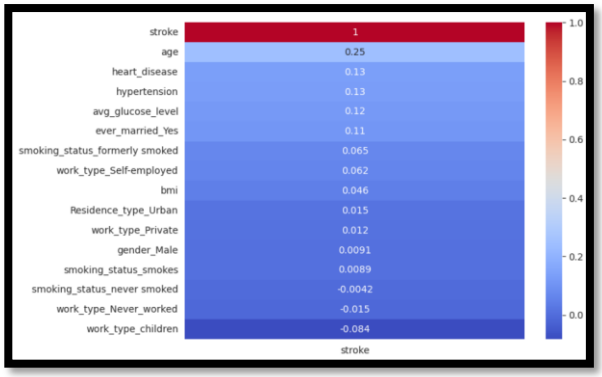


Figure 9 Correlation Matrix

The pandas library's get_dummies function, which transforms categorical data into binary columns that indicate the presence or absence of each category, was used to execute OneHotEncoding. Multicollinearity must not exist for statistical inferences drawn from predictive models to be considered dependable, as this procedure verified it. The correlation matrix of features following OneHotEncoding, as shown in Figure 4.9, indicates that 'stroke' column has the highest correlation value (1), followed by 'age' (0.25). On the other hand, correlation values below 0 for "smoking_status_never_smoked," "work_type_Never_worked," and "work_type_children" indicate little or no impact.

4) Data Balancing

Before applying SMOTE, the dataset contained a total of 5109 samples, with a significant imbalance between positive and negative samples. Specifically, there were 4860 positive samples and a much smaller number of negative samples.

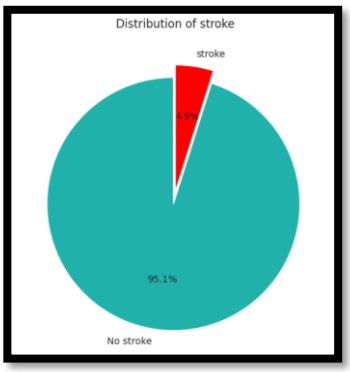


Figure 10 Before using SMOTE

After applying SMOTE, the dataset was balanced, with a total of 9720 samples, evenly split between 4860 positive samples and 4860 negative samples. SMOTE works by generating synthetic examples of the minority class, thereby balancing the dataset without simply duplicating existing minority class instances.

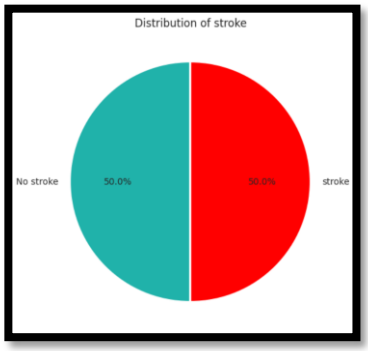


Figure 11 After using SMOTE

5) Data Normalisation

Data normalisation is essential in machine learning, statistics, and data analysis to standardize the scale of features within a dataset. By ensuring that every attribute is within the same range, this technique stops larger features from overshadowing smaller ones and having an unbalanced impact on the learning process. In particular, machine learning algorithms that depend on gradient-based optimization or distance calculations require the use of normalisiation techniques like min-max scaling, which rescales features to a range between 0 and 1. These techniques are essential for enhancing the performance and stability of machine learning algorithms.

Table 1 Min-max Scaler Formula

Performance Evaluation Metrics	Formula

Min-max Scaler	$X\_normalized = \frac{X - X\_minimum}{X\_maximum - X\_minimum}$

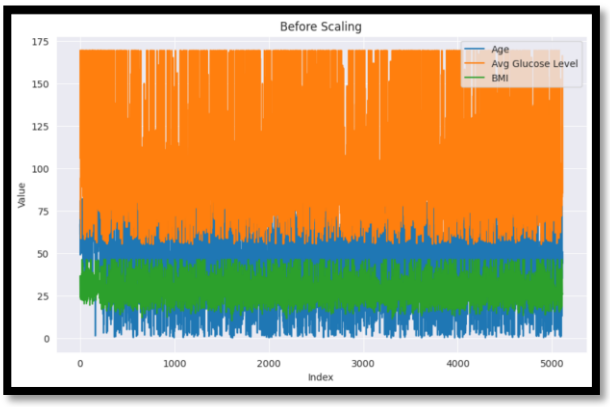


Figure 12 Before Normalisation

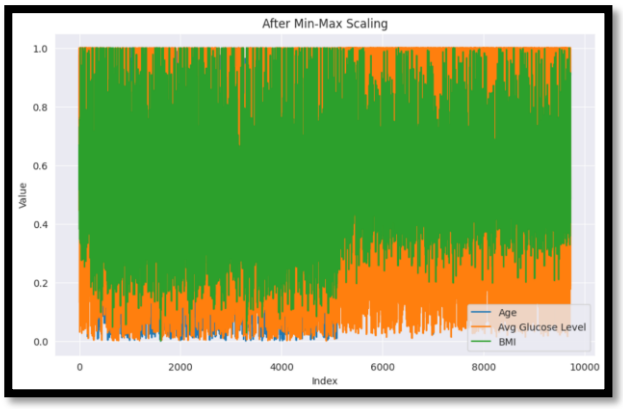


Figure 13 After Normalisation

The features "age," "avg_glucose_level," and "bmi" had different scales and ranges prior to standardization, which could pose problems for some machine learning techniques. By converting these features to a single range between 0 and 1, min-max scaling ensured that each feature contributed equally to the process of training the model. Algorithm stability and performance are enhanced by this normalisation, particularly for algorithms that are sensitive to feature sizes.

6) Feature Selection

A crucial stage in machine learning is feature selection, which finds and keeps the most useful features from a dataset while removing unnecessary or unused ones. By focusing on the most important characteristics, this method increases understanding, decreases overfitting, and boosts model performance. There are several approaches to feature selection, such as embedding techniques like LASSO regression, wrapper techniques like Recursive Feature Elimination (RFE), and filter techniques like SelectKBest. These methods determine which subset of features makes the biggest

contribution to the current predicting task by evaluating each feature's significance using statistical or model performance indicators. According to Guyon and Elisseeff [10], feature selection plays a crucial role in enhancing model generalization and interpretability by reducing the dimensionality of the data while preserving its predictive power.

The SelectKBest method is a feature selection technique commonly employed in machine learning to identify the most relevant features for predictive modeling. It operates by scoring each feature based on its statistical significance in relation to the target variable, typically using metrics like chi-square, ANOVA F-value, or mutual information. SelectKBest then selects the top 'k' features with the highest scores, effectively reducing the dimensionality of the dataset while retaining the most informative features for improving model performance and interpretability.

The Chi-square (χ^2) test is a statistical tool used to determine the existence of a significant association between two categorical variables. It evaluates the difference between observed and expected frequencies of categories under the assumption of independence between the variables. By calculating a test statistic (χ^2), the Chi-square test quantifies the extent of this discrepancy, aiding in determining the strength and significance of the relationship between the variables in question.

Table 2 Chi Square Score for every features

Feature	**Chi-Square Score**
hypertension	1.079137
heart_disease	3.212851
bmi	7.908778
work_type_Never_worked	22.000000
avg_glucose_level	131.158791
smoking_status_never smoked	150.311156
smoking_status_smokes	182.837429
work_type_Private	206.255054
Residence_type_Urban	286.745449
gender_Male	301.023891
ever_married_Yes	340.197318
age	373.697273
work_type_children	602.711485
work_type_Self-employed	624.726761
smoking_status_formerly smoked	701.860825

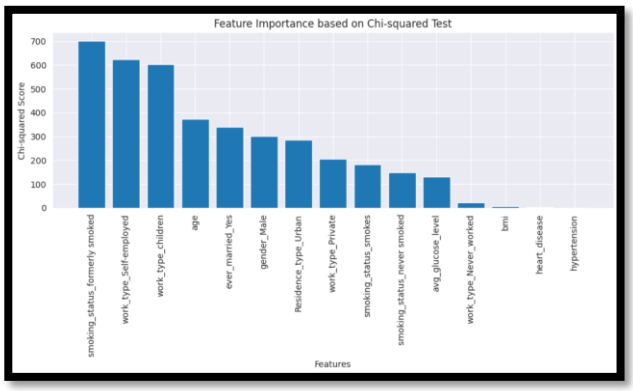


Figure 14 Chi Square Score Bar Plot

Table 2 and Figure 14 shows the Chi-Square scores for the features in the dataset. Features having significantly higher Chi-Square scores (701.86 and 624.73 respectively) in the table and figure include smoking_status (formerly smoked) and work_type (Self-employed), indicating that they may have strong predictive power or significant relationships with the outcome under study. In the provided dataset, features with lower scores such as hypertension and heart disease indicate weaker relationships or less predictive ability.

A crucial phase in the machine learning (ML) process is feature selection, which greatly boosts model performance by lowering complexity, minimizing overfitting, and increasing accuracy. SelectKBest with the Chi-Square test was used in this work to order characteristics according to significance. The top 15 or all features produced the best accuracy, highlighting the significance of feature selection. The comparison of the top 15, top 10, and top 5 feature sets tried to find a balance between accuracy and complexity. By allowing the collection of complex correlations and patterns in the data, offering additional information for improved category discrimination, and reducing underfitting by learning from richer data representation, the selection of more features improves the performance and accuracy of the model.

Table 4 Top Features Accuracy Scores

Top Features	Accuracy
15 (All)	0.8863
10	0.8753
5	0.7751

Table 5 Before and After Feature Selection

Before Feature Selection	After Feature Selection
id	-
stroke	stroke (target variable)
hypertension	hypertension
heart_disease	heart_disease
bmi	bmi
work_type_Never_worked	work_type_Never_worked

avg_glucose_level	avg_glucose_level
smoking_status_never smoked	smoking_status_never smoked
smoking_status_smokes	smoking_status_smokes
work_type_Private	work_type_Private
Residence_type_Urban	Residence_type_Urban
gender_Male	gender_Male
ever_married_Yes	ever_married_Yes
age	age
work_type_children	work_type_children
work_type_Self-employed	work_type_Self-employed
smoking_status_formerly smoked	smoking_status_formerly smoked

D. Data Splitting

Data splitting into 80% training and 20% test sets is fundamental for ensuring model robustness and generalization. The training set serves as a platform for the model to refine its parameters, learning patterns and relationships within the data to minimize errors and make accurate predictions. In contrast, the test set, kept separate from training, evaluates the model's real-world performance by assessing its ability to predict new data. Utilizing the train_test_split function from scikit-learn, the dataset is divided into training and testing subsets, allowing for independent evaluation of the model's performance on unseen data. This separation ensures an accurate assessment of the model's generalization ability, providing insights into its potential efficacy in practical applications.

E. Support Vector Machine Model

Following data pre-processing, the dataset is cleaned and features are chosen for constructing SVM prediction models. These models are built using a loop to determine the time taken for each prediction task. Accuracy is assessed using a confusion matrix, and hyper-parameter tuning optimizes the SVM model's performance. Key hyper-parameters include the regularization parameter (C), kernel type, and kernel coefficient (gamma). The process involves methods like grid search, random search, and Bayesian optimization to find the best settings. In this research, grid search identified optimal parameters: C=10, gamma='scale', and kernel='rbf', similar to Zhang's (2023) study. The model is then trained with these parameters, and predictions are made on both the test and training sets. The SVM model achieved a test accuracy of 89.76%, indicating a high level of performance. Evaluation metrics such as precision, recall, F1-score, and support provide further insights into the model's classification ability..

V. RESEARCH RESULTS

In this section, we examine the results of the experiment described in the previous section, providing insight into the precision and effectiveness of the predictive models created. In addition to presenting the data, we also conduct a thorough analysis and compare our results with those of earlier studies. By comparing our findings to previous research, we want to find trends, identify our advantages and disadvantages, and extract useful information about how well the models we used could predict future events. We want to determine the robustness and reliability of our predictive approach in the field of brain stroke prediction through this extensive study.

A. Performance Evaluation

The confusion matrix forms the cornerstone of evaluating the SVM model's performance, offering a detailed breakdown of its predictions compared to actual stroke classifications in the testing dataset. It delineates true positives, true negatives, false positives, and false negatives. Meanwhile, the Chi-Square Test assesses the significance of the relationship between categorical variables in stroke prediction, selecting features based on their scores to identify those most predictive of stroke presence or absence.

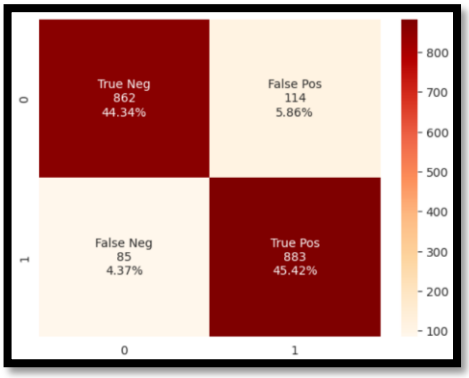


Figure 15 Confusion Matrix

Table 6 Confusion Matrix Values

Confusion Matrix	Value
True Positives	883
True Negatives	862
False Positives	114
False Negatives	85

To fully evaluate the performance of the SVM model, key metrics such as accuracy, precision, recall, and F1 score are also calculated in addition to the confusion matrix. While precision focuses on limiting false positives by properly identifying stroke patients, accuracy measures the total correctness of predictions. The model's recall, also known as sensitivity, measures its ability to identify real-world stroke cases, whereas the F1 score achieves a balance between recall and precision by taking false positives and false negatives into play. This evaluation method offers a thorough understanding of the SVM model's ability to predict for brain strokes.

Table 7 Evaluation Metrics Results

Evaluation Metrics	Results
Accuracy	0.89
Precision	0.89
Recall	0.91
F1 Score	0.90

B. Prior Research Result Comparison

The SVM model's performance was robustly assessed using various metrics, revealing noteworthy findings. Achieving an accuracy of 89.76%, the model demonstrated a high capability in accurately categorizing stroke and non-stroke cases. Additionally, precision, recall, and F1 Score, with scores of 0.89, 0.91, and 0.9 respectively, further underscored the model's effectiveness. A detailed confusion matrix delineated true positives (883), true negatives (862), false positives (114), and false negatives (85), providing insights into the model's strengths and areas for enhancement. Comparatively, a prior study on SVM-based stroke prediction reported an accuracy of 79.2%, suggesting inferior predictive performance compared to the current research. Furthermore, Zhang [8] documented precision, recall, and F1-score values of 0.712, 0.912, and 0.8, respectively, indicating its model's performance characteristics. This research notably outperformed the previous study, underscoring significant improvements in predictive accuracy.

Table 8 Final Result Comparison

Study	Accuracy	Precision	Recall	F1 Score
Current Research	0.89	0.89	0.91	0.90
Zhang, Hanqing. (2023)	0.79	0.71	0.91	0.80

VI. CONCLUSION

This research report demonstrates the development and evaluation of a Support Vector Machine (SVM) model for predicting brain stroke, emphasizing significant advancements in accuracy and reliability compared to previous studies. This study built the SVM model through data preprocessing, Chi-Square-based feature selection, and hyperparameter optimization using a carefully chosen dataset on stroke prediction from Kaggle. The SVM model exceeded earlier benchmarks like Zhang Hanqing's [8] model, which achieved 79.2% accuracy, with an 89.76%

accuracy. The model's strong prediction capacity was demonstrated by a thorough examination using confusion matrices, which showed that it could identify between stroke patients with 91% recall, 89% precision, and an F1 score of 90%.

The research suggests directions for additional research to improve stroke prediction models. These include developing feature engineering approaches to further increase model accuracy, investigating alternative machine learning algorithms including neural networks and ensemble methods, and broadening datasets to cover a variety of demographic and geographic aspects. To improve the model's practical applicability in actual healthcare settings, it is also suggested to collaborate with healthcare providers for clinical validation and to design user-friendly interfaces for smooth integration into healthcare systems.

VII. ACKNOWLEDGMENT

This thesis' preparation was a difficult pursuit. I met many people, academics, and researchers in the fields of computer security and machine learning despite the challenges. They have helped in a variety of ways to finish this research, most notably by expanding my inadequate knowledge in this area of interest.

I would like to express my deepest gratitude and appreciation to the following individuals and organizations who have contributed to the completion of this project. I want to start off by expressing my gratitude to my supervisor Dr. Izyan Izzati Binti Kamsani and my final year project coordinator Dr Nies Hui Wen for their invaluable advice, knowledge, and unwavering support throughout this research project. Their mentoring has been crucial in determining the course of this project.

I also want to thank Universiti Teknologi Malaysia for its academic resources and technical assistance. Without the resources and knowledge offered by the university, as well as the assistance of my dear lecturers, I probably wouldn't be able to finish this thesis. My gratitude and thanks also go out to my friends who provided suggestions or helped with the completion of this thesis, whether indirectly or directly. Their various viewpoints were useful in completing this thesis.

Last but not least, I would also like to thank my family for their unwavering support, encouragement, and understanding throughout this journey. Their love and belief in me have been a constant source of inspiration and motivation. With all the help from my supervisor, friends, lecturers, and family have provided me, I am able to go through all the challenges needed to complete this research.

VIII. REFERENCES

[1] World Health Organization. (2022). WHO EMRO, Stroke, Cerebrovascular accidentHealth topics. World Health Organization - Regional Office for the Eastern Mediterranean.

[2] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., ... & Pellikka, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

[3] Benjamin, E. J., Virani, S. S., Callaway, C. W., et al. (2018). Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation, 137(12), e67-e492.
https://www.ahajournals.org/doi/10.1161/CIR.0000000000000558

[4] Boehme, A. K., Esenwa, C., & Elkind, M. S. V. (2017). Stroke Risk Factors, Genetics, and Prevention. Circulation Research, 120(3),472–495. https://doi.org/10.1161/CIRCRESAHA.116.308398

[5] Xu, Y., Wang, L., He, J., et al. (2019). Prevalence and control of hypertension in Chinese adults. JAMA Network Open, 2(5), e193226. https://pubmed.ncbi.nlm.nih.gov/24002281/

[6] Krishnamurthi, R. V., Ikeda, T., Feigin, V. L., et al. (2018). Global and regional burden of stroke during 1990-2016: findings from the Global Burden of Disease Study 2016. The Lancet Neurology,17(5),439-458. https://pubmed.ncbi.nlm.nih.gov/30871944/

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104 - 116.https://www.sciencedirect.com/science/article/pii/S2001037016300733

[8] Zhang, Hanqing. (2023). Stroke Prediction Based on Support Vector Machine. Highlights in Science, Engineering and Technology. 31. 53-59. 10.54097/hset.v31i.4812.

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

[10] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.