



Early Stroke Disease Prediction Based on Lifestyle Factors Applied with Machine Learning

Suastika Yulia Riska¹, Lia Farokhah²

^{1,2}Informatic Departement, Technology and Desain Faculty, Institut Teknologi dan Bisnis Asia, Malang, Indonesia

¹riska.suastika@asia.ac.id, ²lia.farokhah@gmail.com

Abstract

Stroke prediction has many supporting features and variables. Some forecasts focus more on health or elements that are already present. Predicting stroke risk by identifying habitual factors provides more advantages for preventive action. In addition, the complexity of features or variables is a concern in predicting stroke risk. In this study, we used a public dataset from Kaggle with 10 features or variables. In this study, we propose to collaborate algorithms and preprocessing in feature selection using Pearson Correlation and Principal Component Analysis (PCA) dimension reduction to unravel the complexity of variables and data processing computing. This aims to predict stroke risk more simply. The results of the experiment show that feature selection using Pearson Correlation between features and labels produces maximum results using 5 features out of 10 provided features. This approach produces the best performance on the Naïve Bayes, Iterative Dichotomiser Tree (ID3), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression with 100% accuracy and reduces features by 50% to support the reduction of the complexity of prediction variables and data processing computing.

Keywords: lifestyle; machine learning; preprocessing; stroke; variable reduction

How to Cite: Suastika Yulia Riska and Lia Farokhah, "Early Stroke Disease Prediction Based on Lifestyle Factors Applied with Machine Learning", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 1144 - 1152, Oct. 2025.

Permalink/DOI: <https://doi.org/10.29207/resti.v9i5.6495>

Received: March 19, 2025

Accepted: September 20, 2025

Available Online: October 25, 2025

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Stroke is the leading cause of death and disability worldwide, with significant impacts on public health and healthcare systems [1]. In 2021, stroke was the third leading cause of death globally, down from second place in 2019, and caused about 10% of total deaths. In high-income countries, although most diseases in the top 10 causes of death increased due to population ageing, the number of deaths from stroke was 18%, between 2000 and 2021. However, stroke remains in the top three causes of death [2]. Stroke, which was previously thought to be the dominant disease affecting the elderly group, is now increasingly experienced by younger individuals [3].

This shows that stroke prevention and risk management efforts are crucial in dealing with this significant burden of disease. There are two primary categories of risk factors that contribute to stroke: non-modifiable risk factors, such as age and family history, and modifiable risk factors, such as lifestyle and specific health conditions [1], [4] - [6]. Lifestyle modifications and appropriate health management have the potential to

intervene in modifiable risk factors, which are one of the two categories.

Machine learning (ML) has become a potent instrument in the healthcare industry recently, especially for disease prediction. Large, complicated datasets may be analyzed thanks to machine learning techniques, which can also reveal complex patterns and relationships in the data that are difficult to spot using conventional techniques [4]. In particular, machine learning offers a more reliable and scalable method for determining risk factors in stroke prediction by processing and evaluating various lifestyle and health data, including blood pressure, physical activity, and eating patterns [6].

Previous study has shown that an individual's risk of stroke can be raised by factors such as their diet, physical activity levels, and smoking behaviours [7]. The epidemiological literature also shows that the risk of stroke can be substantially diminished by adopting a healthy lifestyle [8]. Although there is evidence supporting these links, predicting stroke risk remains challenging due to the complexity of lifestyle-related

data. This is where machine learning is critical. By using machine learning methods, we can not only look at big datasets more thoroughly, but we can also learn more about how lifestyle choices affect the risk of having a stroke [1]. The objective of this investigation is to create a classification model that can predict and identify stroke risk by analysing the lifestyle patterns of individuals, utilising machine learning techniques. In this context, the utilisation of machine learning enables a more comprehensive and exhaustive examination of the correlation between a variety of lifestyle factors and stroke risk.

Machine learning, particularly algorithms like Random Forest, SVM, and Logistic Regression, has shown great promise in predicting stroke risks. Random Forest, for example, has been proven to achieve high accuracy rates in stroke prediction models, often surpassing traditional statistical methods. The technique works by building multiple decision trees and combining their results to improve prediction accuracy. Regarding the development of machine learning, specifically Random Forest, it has been demonstrated to achieve superior accuracy in comparison to Decision Tree and Logistic Regression, with an F1 score of 94%. The results of the study applied 10 attributes and showed that BMI is one of the most significant predictor factors [9]. Stroke disease prediction using machine learning using the Random Forest algorithm with an F1 score of 99%. The investigation used 10 attributes without identifying the most crucial factor attribute [10]. Stroke disease classification is carried out by comparing the accuracy of two methods, namely SVM and PSO.

Moreover, by using a Publicly Available Dataset from Kaggle, the SVM approach generated an accuracy of 88% while the PSO-SVM method generated an accuracy of 95%. This model can offer a greater awareness of stroke risk at the individual level by simultaneously examining a variety of lifestyle factors [6]. This enables healthcare providers to offer patients recommendations that are more personalized and precisely targeted. Previous studies on stroke detection has also been conducted using the Naïve Bayes Algorithm and PSO. There are 11 features used, with an accuracy result of 86.21% with the Naïve Bayes Algorithm, and 95.02% with the PSO-Naïve Bayes Algorithm [11]. Research on the early identification system of stroke disease using Reverse Propagation Artificial Neural Network.

The study used 10 features and showed a result of 97.8%. In addition, this study can also provide new insights into the relationship between lifestyle factors and stroke risk [12]. Research on stroke risk prediction using Machine Learning Techniques was also carried out with a stacking method and 10 attributes that showed an accuracy of 98%. By understanding the relative contribution of each factor in influencing stroke risk, more specific and measurable interventions can be developed to prevent or reduce that risk [13]. The Classification and Regression Tree (CART) Algorithm

is also useful for predicting stroke disease. The study used ten features achieving an accuracy of 89.83% [14]. Extreme Gradient Boosting was also used in stroke prediction research, achieving an accuracy of 94.3%. The study included ten attributes [15].

Previous research compared five algorithms for stroke prediction: Random Forest, KNN, Artificial Neural Network, SVM, and Logistic Regression. There are ten features employed, demonstrating that the Random Forest and KNN algorithms outperformed the other three algorithms. The Random Forest has an accuracy of 85%, whereas KNN has an accuracy of 82.5% (16). Another study found that machine learning systems can anticipate strokes with an accuracy of up to 91%. The study used ten features to compare several algorithms. These algorithms are: Logistic Regression, Random Forest, KNN, Naïve Bayes, Support Vector, and xgboost. The Random Forest algorithm [17] provides the maximum accuracy.

The predictive model that will be developed in this study will be able to help healthcare providers provide more personalized recommendations by utilizing machine learning's capacity to analyze the intricate interplay of lifestyle factors. In contrast to conventional methodologies, which may concentrate on a restricted number of variables, this method provides a comprehensive understanding of the correlation between lifestyle choices and stroke risk.

The novel approach employed in this research is the combination of Pearson Correlation for feature selection and PCA for dimensionality reduction, although numerous studies have applied machine learning to stroke prediction. This unique process is important for healthcare applications that focus on being efficient and accurate, as it simplifies the features while keeping their ability to predict outcomes. This study uses a publicly available dataset from Kaggle, which has been widely used in previous research, ensuring the reproducibility of our results. By leveraging this dataset, our study not only builds upon existing work but also offers a new perspective by introducing advanced preprocessing techniques such as PCA and Pearson Correlation. These methods contribute to improving the efficiency of machine learning models in stroke prediction, particularly in terms of computational complexity and feature selection.

The objective of this study is to develop a more practical, scalable, and precise tool for identifying individuals at risk, facilitating timely intervention, and contributing to the global initiative to reduce the prevalence of stroke by integrating machine learning into stroke prediction. Additionally, it is anticipated that this research will facilitate the development of more advanced diagnostic methods and intervention strategies, thereby improving public health outcomes related to stroke prevention.

The results of this study can be used to develop more effective public health policies for encouraging healthy lifestyles and lowering the prevalence of stroke in the community. In the context of globalization and urbanization, it is important to effectively identify and address stroke risk factors [1].

2. Methods

The purpose of this study is to predict stroke early based on the lifestyle of the community. Overall the flow of this research is shown in Figure 1.

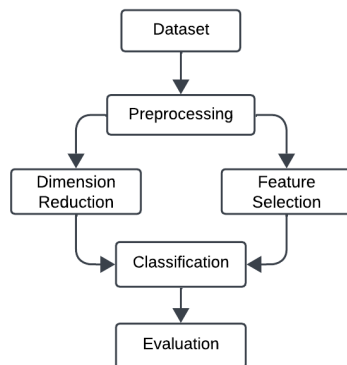


Figure 1. Block diagram of early prediction of stroke disease

Based on Figure 1, the research began with gathering information from which this study used public datasets obtained from Kaggle. Preprocessing—that is, standardising with Z-Score, dimension reduction with PCA, and feature selection with Pearson Correlation—next comes. The classification process follows from there. This work applied five classification techniques: Naives Bayes, Decision Tree ID3, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Logistic Regression. The Confusion Matrix then allowed one to compare the outcomes. Then the results were compared with the Confusion Matrix.

2.1 Dataset

In this study, a public dataset of 4981 data taken from kaggle at the following link <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>. The dataset was divided into training data and testing data with a ratio of 80:20. The dataset consists of several variables that represent the characteristics of the individuals being studied. The Gender variable indicates the respondent's sex, while Age describes the individual's age. The Hypertension variable identifies whether a person has a history of high blood pressure, and Heart_disease shows whether the individual has a heart condition. The Ever_married variable explains the marital status of the respondent, indicating whether they have ever been married or not.

The Work_type variable represents the type of employment, which can include private employees, self-employed individuals, government workers, or children if the data corresponds to younger age groups. The Residence_type variable describes the type of living area, distinguishing between rural and urban

regions. Furthermore, Avg_glucose_level shows the average blood glucose level of an individual, while BMI (Body Mass Index) represents the ratio of weight to height, reflecting body composition. Lastly, the Smoking_status variable describes an individual's smoking habits with categories such as unknown, never smoked, smokes, and formerly smoked.

2.2 Research Scenario

In this study, 3 research scenarios were carried out as shown in Figure 2. Based on Figure 2, there are 3 scenarios that are carried out at the preprocessing stage. Scenario 1 shows all datasets used for all features, which are then classified. In the classification stage, this study applies 5 classification algorithms, namely Naives Bayes, Decision Tree ID3, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression, which are then evaluated with a confusion matrix.

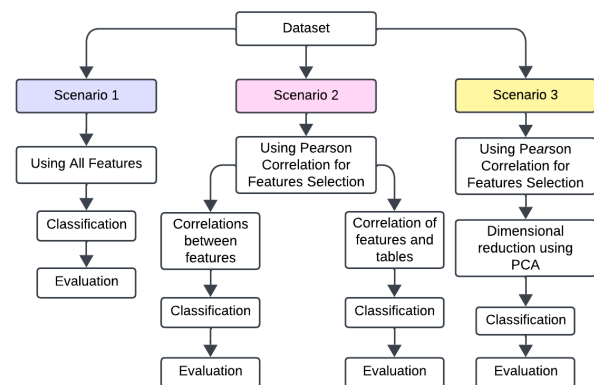


Figure 2. Block diagram of research scenario

In scenario 2, feature selection is carried out using Pearson Correlation. This stage uses two approaches, namely the search for correlations between features and the correlation of features and tables. Then in each approach, classification is carried out with the same algorithm as scenario 1, and evaluated with a confusion matrix. In scenario 3, the same as scenario 2, in the preprocessing stage, feature selection is carried out using Pearson Correlation. However, on scenario 3 This study applied dimensional reduction using Principal Component Analysis (PCA). Furthermore, classification and evaluation are carried out, with the same algorithm on scenario 1 and 2. Several machine learning algorithms are implemented [18], [19].

In this study, the Naive Bayes algorithm was applied to estimate stroke risk by calculating conditional probabilities based on selected lifestyle features. This probabilistic approach was chosen for its simplicity and efficiency in handling independent attributes such as age, hypertension, and glucose level [20]. The Decision Tree C4.5 algorithm was implemented to classify stroke risk by selecting attributes that provided the highest information gain ratio, forming a hierarchical structure that identified the most influential lifestyle factors. This process allowed the model to produce interpretable decision paths that indicate the dominant variables

contributing to stroke occurrence [21]. The Support Vector Machine (SVM) algorithm was utilized to determine the optimal separating hyperplane that distinguishes individuals with potential stroke risk from those without. A kernel function was applied to handle nonlinear relationships among lifestyle variables, improving classification accuracy in high-dimensional feature space [22]. The Logistic Regression model was used to estimate the probability of early stroke occurrence by modeling the relationship between multiple lifestyle predictors and the binary outcome variable. The sigmoid function was employed to map the relationship between input factors and stroke probability, enabling interpretable probability-based classification [23]. The K-Nearest Neighbours (KNN) algorithm was applied to classify new instances based on their similarity to existing data points in the dataset. By calculating the proximity of each individual's lifestyle features—such as BMI, age, and smoking status—to neighbouring samples, the model was able to infer stroke risk based on the closest patterns in the training data. [24], [25].

2.3. Preprocessing

In machine learning, preprocessing is the first stage to prepare raw data such that the algorithm may make most use of it. Raw data unprocessed can include errors, noise, empty data, or improper formats. Preprocessing is therefore necessary to enable the data to become cleaner and more ordered, so enabling the model to operate more precisely and effectively [26]. To mitigate the intricacy of multiple features in stroke prediction, we implemented Pearson Correlation for feature selection and PCA for dimensionality reduction as part of our preprocessing pipeline. Our investigation is a substantial contribution to the field because this combination is rarely investigated in the literature for stroke prediction. Categorical features are modified by label encoding during preprocessing.

Furthermore, the use of a Kaggle dataset guarantees the reproducibility of the results of our study and underscores the potential of public datasets to enhance the generalization and efficacy of machine learning models in healthcare applications. By utilizing publicly available datasets, our research can be independently validated by other researchers, ensuring transparency and fostering further advancements in the field.

The standardization process was performed to transform all numerical features into a common scale with a mean of 0 and a standard deviation of 1. The Z-Score normalization technique was employed to prevent attributes with larger numerical ranges from dominating the computation in machine learning algorithms. This step ensured that all lifestyle-related variables, such as age, BMI, and glucose level, contributed proportionally to the model. [27]. The Z-score equation is shown Formula 1.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

z is standardized data value, x is original value, μ is mean of the data, and σ is standard deviation.

Following data standardization, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while maintaining the maximum amount of information. PCA identifies a set of principal components that represent linear combinations of the original variables with the highest variance. In this research, PCA was used to simplify the dataset and enhance computational efficiency during stroke risk classification, without significantly reducing the predictive capability of the model. The Principal Component Analysis (PCA) process in this study involved several key stages. First, the data were standardized to ensure that each feature contributed equally to the analysis. Next, a covariance matrix was constructed to measure the correlation between the variables. Based on this matrix, eigenvalues and eigenvectors were then calculated, where the eigenvectors represent the directions of the principal components and the eigenvalues indicate the amount of variance explained by each component. Finally, the data were projected onto the principal components with the highest variance, resulting in a lower-dimensional representation that retains most of the essential information. This dimensionality reduction technique is widely applied in various domains, including image processing and genetic data analysis, to simplify data structures while preserving important patterns and relationships [28].

The Pearson Correlation technique was applied in this study to select the most relevant features by quantifying the linear relationship between each variable and the target class. This method calculates the correlation coefficient (r) to determine the strength and direction of association, where values close to +1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values near 0 represent little to no linear relationship. By identifying features with high correlation values to the target variable, less influential or redundant attributes were removed, thereby improving the model's efficiency and interpretability. The Pearson correlation used in this study serves as an essential step in the preprocessing stage to ensure that only meaningful lifestyle-related variables contribute to stroke prediction. The correlation coefficient is mathematically expressed as shown in Formula 2 [29].

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (2)$$

X_i dan Y_i are represent the values of the feature and goal variables, respectively. $\bar{X}$ dan $\bar{Y}$ = represent the means of each variable. Interpretation of Correlation Coefficient, when $r < 0$ is Positive correlation (if X increases, Y also increases). $r < 0$ is Negative correlation (if X increases, Y decreases). $r = 0$ is No linear correlation.

2.4 Evaluation Using the Confusion Matrix

In this study, the evaluation used to measure the performance of the algorithm is Confusion Matrix, which includes accuracy equation is shown in Formula 3, precision equation is shown in Formula 4, recall equation is shown in Formula 5, and F1-score equation is shown in Formula 6.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1 Score} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision}+\text{Recall}} \quad (6)$$

TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative [30]. The way the Confusion Matrix works is by comparing the number of predictions with correct values and the number of predictions with wrong values in the prediction class whether or not they have a stroke. The description of the Confusion Matrix is shown in Table 3.

Table 3. Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

3. Results and Discussions

3.1 Preprocessing Features

This study will compare the performance of feature selection with two different methods, namely feature selection with a dimension reduction approach and feature selection using correlation techniques. The

difference between these two approaches is that Principal Component Analysis (PCA) is used to reduce dimensions by turning the original feature into a main component without reducing the feature whereas Pearson Correlation selects features that have a strong correlation with the target, eliminating features that have a low or insignificant correlation. In the preprocessing process, three things are carried out, namely standardization using Z-Score, dimension reduction using PCA, and feature selection using pearson correlation. Using the PCA approach, the dimensions are reduced from ten to three major features. Figure 3 depicts an overview of the PCA-based feature transformation findings.

	PC1	PC2	PC3
0	2.545247	3.652845	1.178087
1	2.282765	2.303843	-0.837016
2	1.634646	-0.632934	0.958564
3	1.700947	1.529020	-1.145725
4	1.469942	1.155522	1.025082
...	...	...	...
4976	-0.927145	-0.003126	-0.993150
4977	1.288481	0.476637	0.988576
4978	2.237912	-0.583123	-1.229674
4979	0.694555	-0.545549	-1.020161
4980	2.075185	0.100636	0.814883

Figure 3. Visualization of PCA methods

PCA reduction techniques help reduce the number of features (variables) in a dataset without losing significant information. In this study, this aims to reduce the dimensions of quite a lot of variables, namely 10 variables and improve their performance.

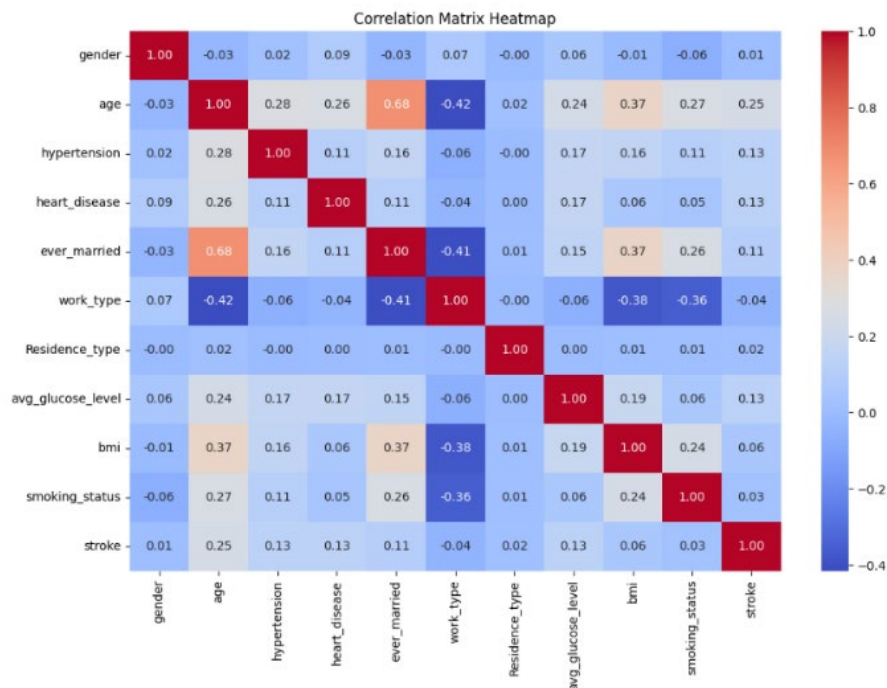


Figure 4. Pearson correlation results

At the feature selection stage using pearson correlation, there are two approaches. The first approach is to look for correlations between features and the second is the correlation of features and labels. In the first approach, the "ever_married" feature is reduced while leaving 9 features, namely, gender, age, hypertension, heart_disease, work_type, Residence_type, avg_glucose_level, bmi, smoking_status. In the second approach that is linked to the label, there are 5 features that are maintained age, heart_disease, avg_glucose_level, hypertension, ever_married. An overview of the correlation results obtained with the implementation in python is shown in Figure 4.

3.2 Performance Using All Features

In the first stage, the performance of 5 machine learning algorithms was evaluated to see the results of stroke prediction with 10 features or variables. The performance results were evaluated using a confusion matrix. The performance results are shown in Table 4.

Table 4. Performance evaluation results with all features

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	1.00	1.00	1.00	1.00
Decision Tree ID3	1.00	1.00	1.00	1.00
SVM	1.00	1.00	1.00	1.00
KNN	0.94	0.91	0.94	0.92
Logistic Regression	1.00	1.00	1.00	1.00

In Table 4, it shows that four machine learning algorithms are very resilient in completing stroke disease classification so that they get 100% accuracy results, namely the bayes naives algorithm, ID3, Support Vector Machine (SVM), and logistic regression. The KNN algorithm gets a lower accuracy of about 94%. The performance shown has computational weaknesses and complexity of features or variables where it consists of 10 features so that further research is carried out to reduce features and select features using PCA and Pearson Correlation.

3.3 Performance Evaluation Involving PCA

At the performance stage, it uses the Z-score and PCA standardization approach. There are 5 algorithms used to test stroke classification performance. The performance results of each algorithm are compared in Table 5.

Table 5. Results of performance evaluation involving PCA

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	0.94	0.91	0.94	0.92
Decision Tree ID3	0.90	0.91	0.90	0.91
SVM	0.95	0.89	0.95	0.92
KNN	0.94	0.89	0.94	0.92
Logistic Regression	0.95	0.89	0.95	0.92

From the results of Table 5, it is known that the collaboration of PCA dimension reduction and SVM machine learning algorithms and Logistic regression has high performance with the same accuracy of 95%,

followed by the Naives Bayes, KNN, and Decision tree ID3 algorithms have the lowest performance. Dimensional reduction using PCA results in more degraded performance than when using all features without reduction as shown in Table 4.

3.4 Performance Using Pearson Correlation

Feature selection using pearson correlation is divided into two, namely correlation between features or variables and correlation between features and labels. In the correlation between features, one feature reduction is produced, namely "ever_married". The results of the algorithm performance by omitting one feature, namely "ever_married" are shown in Table 6.

Table 6. Pearson Correlation Performance Results By Reducing A Single Feature

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	1.00	1.00	1.00	1.00
Decision Tree ID3	1.00	1.00	1.00	1.00
SVM	1.00	1.00	1.00	1.00
KNN	0.94	0.91	0.94	0.92
Logistic Regression	1.00	1.00	1.00	1.00

In Table 6, when reducing data using the Pearson correlation between variables, the performance tends to be good for all algorithms, especially the Naives Bayes, ID3, SVM, and Logistic Regression algorithms. Next is a feature selection approach using Pearson Correlation that connects features and labels or outputs. The results are shown in Table 7.

Table 7. Results Of Performance Evaluation Involving Pearson Correlation Label

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	1.00	1.00	1.00	1.00
Decision Tree ID3	1.00	1.00	1.00	1.00
SVM	1.00	1.00	1.00	1.00
KNN	0.94	0.92	0.94	0.92
Logistic Regression	1.00	1.00	1.00	1.00

In Table 7, Feature reduction using the correlation feature with a threshold value with a label of 0.1 produces 5 features with very satisfactory performance results, namely hypertension, heart_disease, ever_married, age, avg_glucose_level features. In addition to producing satisfactory performance, namely with 100% accuracy on Naives bayes, ID3, SVM, Logistic Regression algorithms. Each of these features plays a crucial role in determining stroke risk. Hypertension is a well-established medical risk factor, as elevated blood pressure can damage blood vessel walls and increase the likelihood of clot formation or rupture in the brain, thereby strongly enhancing the model's predictive capability. Heart disease also contributes significantly, since cardiovascular conditions such as heart failure or atrial fibrillation can cause emboli formation, which obstructs cerebral blood flow and raises the risk of ischemic stroke. The ever_married feature reflects lifestyle and behavioral

variations related to marital status, where differences in diet, stress, and physical activity between married and unmarried individuals may influence stroke susceptibility. Age remains a fundamental non-modifiable risk factor, as advancing age brings physiological changes such as reduced arterial elasticity and a higher incidence of comorbidities like hypertension and diabetes, increasing stroke vulnerability. Lastly, the average glucose level feature is strongly linked to diabetes, which can lead to vascular damage and atherosclerosis due to chronic hyperglycemia, thereby elevating the risk of cerebrovascular events. Collectively, these five features provide an interpretable and biologically meaningful foundation for early stroke prediction based on lifestyle and health indicators. Overall, these five features are not only statistically significant within the model but are also supported by strong clinical evidence. The combination of clinical factors (hypertension, heart disease, glucose level, and age) with a sociodemographic factor (marital status) provides a more comprehensive view of stroke risk, thereby enhancing both the accuracy and clinical relevance of the developed predictive system.

3.5 Performance Evaluation Involving Pearson Correlation Between Variables And PCA

At this stage, feature selection is carried out using Pearson correlation and PCA dimension reduction. Features are selected first with Pearson correlation and then reduced with PCA. The results of the Pearson correlation collaboration between variables, the PCA reduction algorithm, and the algorithm are shown in Table 8.

Table 8. The Results of The Performance Evaluation Involving Pearson Correlation Between Variables And PCA

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	0.94	0.92	0.94	0.92
Decision Tree	0.91	0.91	0.91	0.91
ID3				
SVM	0.95	0.89	0.95	0.92
KNN	0.93	0.90	0.93	0.92
Logistic	0.95	0.95	0.95	0.92
Regression				

Table 9. The Results of The Performance Evaluation Involving Pearson Correlation Label And PCA

Algorithm	Confusion Matrix Evaluation Performance			
	Accuracy	Precision	Recall	F1 score
Naives Bayes	0.91	0.93	0.91	0.92
Decision	0.91	0.90	0.91	0.90
Tree ID3				
SVM	0.95	0.89	0.95	0.92
KNN	0.94	0.90	0.94	0.92
Logistic	0.95	0.89	0.95	0.92
Regression				

From the results of Table 8, there is a decrease in accuracy in all algorithms so this method is not recommended. The next experiment involves the

selection of Pearson Correlation features between features and labels by involving PCA shown in Table 9.

In Table 9, there is also a decrease in all algorithms compared to only using the selection of the Pearson correlation feature without dimension reduction.

3.5 Discussions

Unlike previous studies that primarily focused on maximizing classification accuracy, this study places greater emphasis on reducing the number of features through techniques like Pearson Correlation and Principal Component Analysis (PCA). By identifying and retaining only the most relevant features, this approach not only enhances computational efficiency but also provides more interpretable and actionable insights. The significance of feature selection in the prediction of stroke risk is underscored by the results of this study, particularly in the reduction of computational costs and feature complexity while preserving high accuracy.

These results are compared to those of previous research, and several significant observations are revealed. For example, Random Forest has been employed in studies to achieve F1 scores as high as 99% [10], despite the absence of identification of the most critical features. Other studies that employed SVM and PSO-SVM achieved accuracy rates of 88% and 95%, respectively, which further substantiated the significance of SVM in stroke classification [6]. In a similar vein, research employing Naïve Bayes and PSO-Naïve Bayes achieved accuracy rates of 86.21% and 95.02%, respectively, while Reverse Propagation Artificial Neural Networks achieved 97.8% accuracy [12].

The present research prioritizes feature selection as a critical step in improving efficiency, in contrast to previous studies that primarily focused on optimizing classification accuracy using full feature sets. By reducing the number of features by 50%, the model effectively reduces computational complexity while maintaining a high level of accuracy across multiple algorithms, particularly SVM and Logistic Regression. This approach aligns with prior research showing that SVM consistently demonstrates superior performance in stroke classification tasks. Furthermore, this study identified five key features that significantly contribute to stroke risk: age, heart disease, average glucose level, hypertension, and ever married. These findings suggest that lifestyle interventions, particularly those focusing on glucose intake and hypertension management, could be particularly effective in reducing stroke risk. By emphasizing the most influential habitual factors, this research provides not only a model with high accuracy but also one that can inform preventive measures in public health.

A key contribution of this study is its exploration of the relative contribution of each feature to stroke risk, an area that has not been explicitly examined in many prior

studies. By isolating and focusing on specific lifestyle factors, such as hypertension control and glucose intake, this study offers a practical perspective on stroke prevention strategies. This differs from previous research that mainly concentrated on optimizing model performance metrics. In comparison to studies that have concentrated on full-feature sets to improve accuracy, this study highlights the importance of feature selection as a critical step to improve model efficiency. The reduction in the number of features allowed the model to maintain high accuracy while simultaneously reducing computational complexity. This trade-off between accuracy and efficiency is crucial in real-world healthcare scenarios, where speed and resource optimization are just as important as predictive power.

Overall, the findings underscore that the optimization of stroke prediction models is significantly influenced by feature selection, particularly through Pearson Correlation. This method facilitates the practical implementation of machine learning-based stroke prediction systems in real-world healthcare applications by reducing feature complexity and enhancing computational efficiency. Additionally, by focusing on the relative contribution of lifestyle factors, this study provides actionable insights that can inform healthcare policies, such as encouraging dietary changes and lifestyle adjustments to reduce stroke risk.

4. Conclusions

This research focuses on feature reduction to reduce feature complexity and later can reduce computation in data processing. In research, stroke risk prediction through more features in habits can be one of the breakthroughs for early prevention of stroke risk. The dataset used is taken from the public dataset in Kaggle. Three aspects have been focused on: improving accuracy performance. Second, using the Pearson correlation feature selection method, both correlation between variables and correlation with labels. Third, collaboration in feature selection and dimension reduction is carried out. The purpose of this selection is to reduce the complexity of features or variables and will reduce the computation of data processing. From the results of pre-processing and collaboration experiments, stable SVM and Logistic regression algorithms were obtained in various treatments. This can be seen in the comparison of performance in Table 4- Table 9 and Table 7 as the best results. The identification of five key features : hypertension, heart_disease, ever_married, age, and avg_glucose_level highlights both their statistical significance and clinical relevance. Among these, reducing glucose intake and managing hypertension-triggering foods in daily life play an important role in minimizing stroke risk. In conclusion, this study offers a novel approach to stroke risk prediction by integrating PCA for dimensionality reduction and Pearson Correlation for feature selection, a combination that is not commonly found in previous stroke prediction

studies. Additionally, the use of a publicly available Kaggle dataset adds value by ensuring the reproducibility of our findings, which is essential for advancing machine learning applications in healthcare and promoting their wider adoption in predictive modeling.

However, there are some limitations to this study. Future research should incorporate cross-validation or stratification, particularly in cases with imbalanced classes, to determine whether the accuracy claims presented in this study reflect overfitting. Additionally, we specifically highlight that the use of public datasets from Kaggle, while clinical in nature, may not fully reflect the complexity of real-world patient data, particularly due to differences in population characteristics across regions. We also acknowledge that the developed model needs further testing on larger-scale and more diverse datasets to ensure scalability and the generalizability of the results.

As a direction for future research, although this study has utilized real clinical data from the publicly available Kaggle dataset, further validation with local clinical data from each country is still required. This is important because population characteristics, lifestyle factors, and stroke risk patterns may vary across regions. Consequently, the developed prediction model can become more accurate, relevant, and ready for practical implementation in specific clinical contexts.

Acknowledgements

Thank you to LPPM and the ASIA Malang Institute of Technology and Business for supporting this internal research so that the research can be completed properly.

References

- [1] V. L. Feigin, "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990-2019: A Systematic Analysis for The Global Burden of Disease Study 2019," *Lancet Neurol.*, Vol. 20, No. 10, Pp. 1–26, 2021, DOI: 10.1016/S1474-4422(21)00252-0.
- [2] World Health Organization, "WHO - The Top 10 Causes of Death," 24 May, 2024. <http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death>.
- [3] I. Gusti Nyoman Cahya Aditya, G. Angga Pradipta, I. Gusti Agung Vony Purnama, and E. Sulistyono Rini, "Introduction to Stroke Disease Through Website-Based Interactive Multimedia," *Spinter*, Vol. 1, No. 2, Pp 363-368, 2024.
- [4] H. Lee et al., "Machine Learning Approach to Identify Stroke Within 4.5 Hours," *Stroke*, Vol. 51, No. 3, pp. 860–866, 2020, DOI: 10.1161/PP.119.027611.
- [5] H. Mustamu et al., "Post-Stroke Rehabilitation Center In Merauke : Healing Environment Architectural Design Approach," *Musamus Journal of Architecture*, Vol. 6, No. 1, Pp. 1–10, 2023. DOI: <https://doi.org/10.35724/mja.v6i1.5893>.
- [6] Y. Ayuningtyas and I. Made Suartana, "Classification of Stroke Diseases Using Support Vector Machine (SVM) and Particle Swarm Optimization (PSO)," *JINACS (Journal Informatics Comput. Sci.)*, Vol. 04, No. 04, Pp. 452–457, 2023. DOI : <https://doi.org/10.26740/jinacs.v4n04.p451-457>
- [7] M. Rahmadhani, "Factors Affecting the Occurrence of Hypertension in the Community in Bedagai Village, Pinang City," *J. Dock. STM (Science and Technology. Med.)*, Vol. 4, No. 1, Pp. 52–62, 2021, DOI: 10.30743/stm.v4i1.132.
- [8] S. Hoshida et al., "Morning and Evening Home Blood Pressure

- and Risks of Incident Stroke and Coronary Artery Disease in The Japanese General Practice Population: The Japan Morning Surge-Home Blood Pressure Study," American Heart Association. Hypertension, Vol. 68, No. 1, Pp. 54–61, 2016, DOI: 10.1161/USSR.116.07201.
- [9] O. Shobayo, O. Zachariah, M. O. Odusami, and B. Ogunleye, "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm," Analytics, Vol. 2, No. 3, Pp. 604–617, 2023, DOI:10.3390/analytics2030034.
- [10] M. Putri, "Prediction of Stroke Disease Using Machine Learning with Random Forest Algorithm," J. Infomedia Tek. Inform., Vol. 9, No. 2, Pp. 16–21, 2024. DOI: 10.30811/jim.v9i1.5199
- [11] M. S. Hasibuan, D. Fransisca, J. Magister, T. Informatics, and F. I. Computer, "The Use of Naive Bayes Algorithm and Particle Swarm Optimization (PSO) to Detect Stroke," Pp. 109–118, 2022. DOI: <https://doi.org/10.31284/j.integer.0.v9i1.5738>
- [12] L. S. A. Putra, Eka Kusumawardhani, Putranty Widha Nugraheni, Lalak Tarbiyatun Nasyin Maleiva, and Vincentius Abdi Gunawan, "Early Identification System of Stroke Disease Using Reverse Propagation Artificial Neural Network," J. Technol. Inf. J. Science and Study. Bid. Tech. Inform., Vol. 16, No. 2, Pp. 145–157, 2022, DOI: 10.47111/jti.v16i2.5096.
- [13] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors, Vol. 22, No. 13, 2022, DOI: 10.3390/s22134670.
- [14] A. Fadillah Hermawan, F. Rakhmat Umbara, and F. Kasyidi, "MIND (Multimedia Artificial Intelligent Networking Database Early Prediction of Stroke Disease Based on Medical Records using the CART (Classification and Regression Tree) Algorithm Method," J. MIND J. | ISSN, Vol. 7, No. 2, Pp. 151–164, 2022, [Online]. DOI: <https://doi.org/10.26760/mindjournal.v7i2.151-164>.
- [15] D. T. Muriandisyah, "Stroke Prediction Using Extreme Gradient Boosting," JIKO (Journal of Inform. and Computers), Vol. 8, No. 2, Pp. 419, 2024, yogurt: 10.26798/geco.v8e2.1295.
- [16] W. W. Liao, Y. W. Hsieh, T. H. Lee, C. ling Chen, and C. yi Wu, "Machine Learning Predicts Clinically Significant Health Related Quality of Life Improvement After Sensorimotor Rehabilitation Interventions in Chronic Stroke," Sci. Rep., Vol. 12, No. 1, Pp. 1–10, 2022, DOI: 10.1038/s41598-022-14986-1.
- [17] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," IEEE Access, Vol. 11, No. May, Pp. 52288–52308, 2023, DOI: 10.1109/ACCESS.2023.3278273.
- [18] I. H. R. Hatta et al., Artificial Intelligence. Cendikia Mulia Mandiri, 2024.
- [19] I. D. Id, Machine Learning: Theory, Case Studies and Implementation Using Python, Vol. 1. Unri Press, 2021.
- [20] L. M. A. Arrozi, A. Y. Rahman, and R. P. Putra, "Klasifikasi Tingkat Tutor Bahasa Sasak Berbasis Teks Menggunakan Naive Bayes." JITET (Jurnal Informatika dan Teknik Elektro Terapan), Vol.12, No.3, Pp 3556-3565, 2024, DOI: <http://dx.doi.org/10.23960/jitet.v12i3S1.5149>
- [21] M. Maulidah, W. Gata, R. Aulianita, and C. I. Agustyaningrum, "Decision Tree Classification Algorithm for Book Recommendations Based on Book Category," E-Business J. Ilm. Economist. and Business, Vol. 13, No. 2, Pp. 89–96, 2020. DOI: <https://doi.org/10.51903/e-bisnis.v13i2.251>
- [22] I. P. D. W. Darmawan, G. A. Pradnyana, and I. B. N. Pascima, "Optimization of Support Vector Machine Parameters with Genetic Algorithm for Sentiment Analysis on Instagram Social Media," SINTECH (Science Inf. Technol. J., Vol. 6, No. 1, Pp. 58–67, 2023. DOI: <https://doi.org/10.31598/sintechjournal.v6i1.1245>
- [23] M. Saleh, "Classification of Student Satisfaction of MTS Surya Buana Using The Logistic Regression Method." Maulana Malik Ibrahim State Islamic University, 2024.
- [24] L. Farokhah and P. Correspondence, "Implementation of K-Nearest Neighbor For Flower Classification With Extraction of RGB Color Features", DOI: 10.25126/jtiik.202072608.
- [25] S. Y. Riska and L. Farokhah, "Classification of Indonesian Kitchen Spices Using the K-Nearest Neighbors (K-NN) Method," SMATIKA Jurnal. Vol. 11, No. 01, 2021. DOI : <https://doi.org/10.32664/smatika.v11i01.568>
- [26] Xiao, Ziwei. Gang, Wenjie Yuan et al, " Impacts of data preprocessing and selection on energy consumption prediction model of HVAC systems based on deep learning," Energy and Buildings., Vol. 258, 1 March , 2022. DOI: <https://doi.org/10.1016/j.enbuild.2022.111832>
- [27] A. R. Piandy, "Heart Attack Detection Based on Medical Record Data Using The Support Vector Machine." Maulana Malik Ibrahim State Islamic University, Proceedings The 2nd Annual Dharmawangsa International Conference. 2024.
- [28] I. Werdiningsih, M. Kom, B. Nuqoba, M. Kom, and S. S. Muhammadun, Data Mining Using Android, Weka, and SPSS. Airlangga University Press, 2020.
- [29] V. A. Windarni, A. F. Nugraha, S. T. A. Ramadhani, D. A. Istiqomah, F. M. Puri, and A. Setiawan, "Detection of Phishing Websites Using Filter Techniques on Machine Learning Models," Inf. Syst. J., Vol. 6, No. 01, Pp. 39–43, 2023. DOI: <https://doi.org/10.24076/infosjournal.2023v6i01.1268>
- [30] Lia Farokhah and Suastika Yulia Riska, "Analysis and Development of Eight Deep Learning Architectures for the Classification of Mushrooms," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), Vol. 8, No. 1, Pp. 142–149, Feb. 2024, DOI: 10.29207/resti.v8i1.5498.