Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 8, No. 4, November, Pp. 729-740

Enhancing accuracy on chronic-kidney disease detection using machine learning with technique of resampling and missing value treatment

Muhammad Raihan Wibowo[1], Irma Palupi[*1]
School of Computing, Telkom University, Indonesia[1]

Article Info
Keywords:
Chronic Kidney Disease, Support Vector Machine, Logistic Regression, Principal Component Analysis, SMOTE

Article history:
Received: June 23, 2023
Accepted: September 04, 2023
Published: November 30, 2023

Cite:
M. R. Wibowo and I. Palupi, "Enhancing Accuracy on Chronic-Kidney Disease Detection Using Machine Learning with Technique of Resampling and Missing Value Treatment", KINETIK, vol. 8, no. 4`, Nov. 2023.
https://doi.org/10.22219/kinetik.v8i4.1761

*Corresponding author.
Irma Palupi
E-mail address:
irmapalupi@telkomuniversity.ac.id

Abstract
Chronic kidney disease is one of the deadliest diseases in the world. It is important to identify chronic kidney disease at an early stage, so that treatment and prevention can be carried out early. This study used linear interpolation method to treat the missing values, resampling using SMOTE method, and several feature selection methods, such as Pearson's correlation coefficient and Principal component analysis. For the classification methods, Support Vector Machine and Logistic Regression were used to build prediction models for chronic kidney disease based on dataset on UCI Machine Learning. To measure the performance of the model, several test scenarios were tested out so it can be compared to the previous research on the detection of chronic kidney disease, which is used as a benchmark for this study. The best result from the experiment is obtained from the scenario of resampling using SMOTE and feature selection using Principal Component Analysis with averaged accuracy, precision, and f1-score respectively are 98,8%, 100%, dan 98,77%.

1. Introduction

Chronic Kidney Disease (CKD) is a type of kidney disease that causes gradual decline in kidney function. This phenomenon can be observed over months or even years due to varying patient lifestyles. CKD is also known as renal failure, and according to current medical statistics, 10% of the world's population suffers from chronic kidney disease. In 2005, approximately 58 million people died, and according to the World Health Organization (WHO), 35 million of those deaths were related to chronic disease [1]. The diagnosis of CKD typically begins with clinical data, laboratory tests, imaging studies, and ultimately, biopsy. Although biopsy is a standard diagnostic test, it has several drawbacks, such as being invasive, expensive, time-consuming, and occasionally risky. For instance, if a biopsy is performed, patients may experience facial swelling, fear of surgery, and potential misdiagnosis. Imaging techniques (such as mammography, sonography, and renal MRI) have been used to detect this disease for many years. However, there are limitations to their use, including concerns about radiation exposure. Despite these risks, the information obtained from imaging is not sufficient for diagnosing CKD [1].

A data science solution for the analysis of healthcare data is prospective to help saving lives and improve the quality of life. Data is one of the most crucial resources for researchers in the fields of health sciences and medicine. Through those sample data, analysts search for trends, patterns, and similarities to invent new treatments or strengthen existing ones. Providing sample data may be expensive, and rare strains (or fewer samples) may not appear to have enough information to make the prediction statistically significant. Thus making the learning process from the dataset more challenging to use appropriately [2]. Predictive analysis is a method that utilizes various techniques, such as machine learning, data mining, and statistics to forecast future events. In the fields of health care and medicine, this technique has benefits for data analysis in order to evaluate, make decisions, and make predictions [3]. The goal of predictive analysis is to transform data into valuable insights that can enhance businesses [4]. In the context of classifying kidney failure, one can create a classifier using predictive and classification methods such as support vector machines (SVM) and logistic regression (LR).

Research [5] discusses how kidney disease has been a major concern in the healthcare industry. The healthcare industry generates a large amount of data that needs to be examined to uncover hidden information for diagnosis, prognosis, and decision-making. Kidney failure is currently one of the leading causes of death in India. Kidney failure is the gradual loss of kidney function. As the disease worsens, it leads to problems such as high blood pressure, anemia, weak bones, malnutrition, and nerve damage. The Global Burden of Disease (GBD) 2015 ranked kidney failure as the eighth leading cause of death in India.

Predictive analysis predicts future events that have not yet occurred or predicts conditions that have not been observed from the historical data. By examining historical data, patterns can be identified, which can then be used to establish relationships or map these patterns to future events. Predictive analysis involves the use of statistical analysis and machine learning techniques to determine the likelihood of future events or conditions not present in previous data. The models used generate probability predictions for different or new data [6].

This research used references from previous studies on methods and models for analyzing and predicting kidney disease events based on the medical histories of patients. The inquiry utilizes several sources. The first study, conducted in 2018 by Amirgaliyev et al.[1], utilized machine learning and support vector machines (SVM) to analyze chronic renal disease with an accuracy of 82%. The author intended to improve accuracy by employing SVM with feature selection. The research has concluded that the method Support Vector Machine (SVM) has an accuracy of 82%. In this study, the writer hopes that SVM with feature selection can have better accuracy. In 2020, I. U. Ekanayake [7] conducted research on chronic kidney disease, Little's MCAR (Missing Completely at Random) mechanism was employed, a statistical technique designed to manage the complexities introduced by these absent values in a systematic manner, and the prediction method used was support vector machine, which achieved an accuracy of 96%. The writer wants to explore whether using linear interpolation for handling missing values will yield better accuracy. In 2023, S. Pal [8] conducted a research on a chronic renal disease dataset using SVM, examining three variations: categorical features, non-categorical features, and a combination of both. The respective attained accuracies were 91%, 91%, and 88%. The purpose of the present study is to examine SVM with two feature selection methods—correlation matrix and principal component analysis (PCA) in order to evaluate their effect on accuracy. D.A. Debal [9] utilized SVM with recursive feature elimination and unsupervised feature selection in 2022, attaining 95.5% and 96.6% accuracy, respectively. The purpose of the present investigation is to examine whether PCA is more accurate than previous research. In 2021, G.M. Ifraz [10] focused on chronic kidney disease and removed outliers using correlation-based feature selection. The accuracy of the logistic regression algorithm was 97%. Ongoing research investigates linear interpolation and resampling synthetic minority oversampling technique (SMOTE) as methods for managing missing data in the dataset on chronic kidney disease. In the second study conducted in 2021, Emon et al. [11] used principal component analysis and logistic regression to predict chronic kidney disease with an accuracy of 96%. With or without resampling, the author anticipates that PCA could conceivably improve accuracy.

On the chronic kidney disease dataset, this study aims to attain optimal performance using classification methods, namely Support Vector Machine (SVM) and Logistic Regression (LR). The newness comes from the use of a few techniques: linear interpolation is used to fill in missing values, correlated features and principal component analysis (PCA) are used to choose features, and the Synthetic Minority Oversampling Technique (SMOTE) is used to fix imbalanced data. Not only do these techniques function as classifiers, but they also model class probabilities, resulting in remarkable accuracies of 98.8% and 98.6% respectively and the runtime model. The unique contribution of this study is its exhaustive examination of the combined effect of improved missing value treatment, resampling techniques for unbalanced datasets, and advanced feature selection methods on the accuracy of prediction systems for machine learning models.

2. Research Method

2.1 System Design

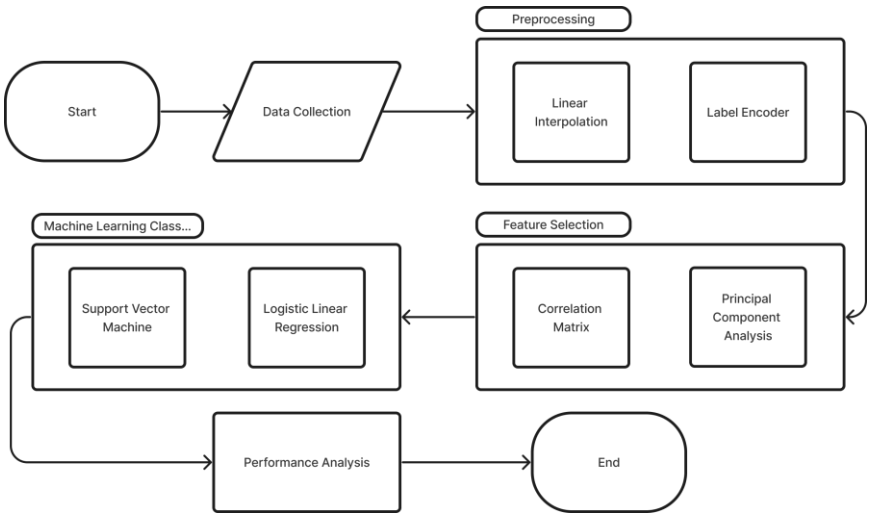

Figure 1. System Design

Figure 1 shows the comprehensive system design of this research. This design encapsulates a multi-faceted approach, incorporating various elements that synergistically contribute to the program's functionality and objectives. This study refers to several previous research references regarding methods and models for the analysis and prediction/detection of kidney disease events in patients based on their medical history.

2.2 Data Collection

The kidney disease dataset is obtained from the UCI Machine Learning website. The dataset chosen is the kidney disease dataset in India for the year 2015. It contains 24 features for 400 individuals [12], including 14 numerical features and 10 categorical features. Several numerical features exhibit extreme outliers, while others have discrete values. The data becomes discrete due to the measurement methods. Some features have a high proportion of certain values, making them not easily associated with measures of central tendency. Some features are highly skewed, while others are normally distributed. While certain categorical characteristics have nearly no missing values, others have a very high fraction of them. However, in this research the models are built from non-categorical data only.

2.3 Preprocessing

2.3.1 Interpolating Missing Value

Missing values in the dataset are smartly filled using a localized approach called linear interpolation. It involves estimating missing values using nearby data points. The algorithm looks at the data points just before and after the missing one to establish a straight-line connection between them. This relationship is then used to predict the missing value, replacing it with a value that fits the existing pattern. The advantage of using local linear interpolation is that it generates estimations that suit the dataset's specific characteristics. It focuses on the immediate data surroundings and captures likely trends and patterns nearby [13]. This leads to more accurate estimations of missing values compared to methods that consider the entire dataset and might miss small local differences. By using linear interpolation to handle missing values, the dataset's accuracy is maintained, benefiting any future analysis or modeling with a completer and more consistent dataset. This approach is especially valuable when missing values sporadically arise due to the influence of neighboring data points. It enables accurate guesswork based on data while keeping the dataset's natural structure intact.

The implementation is just as usual linear interpolation. Suppose that (x_1, y_1) and (x_2, y_2) be the adjacent data points near point (x, y), where y is the missing value [14]. Then the guess value for y can be calculated using Equation 1.

$$\hat{y} = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{1}$$

2.3.2 Resampling using SMOTE

The SMOTE technique is employed as a strategic approach in analyzing an imbalanced dataset related to kidney disease from the UCI repositor. The dataset's inherent imbalance, where one class significantly outnumbers the other, can lead to skewed model performance. By applying SMOTE, synthetic instances are generated for the minority class through interpolation of existing data points. A sample of the minority class is chosen, and fresh synthetic samples are added along the line segments connecting some or all of the minority class's k-nearest neighbors [15]. It effectively balances the class distribution, ensuring that the model is exposed to a more representative set of data. So, the machine learning model can learn from a more complete and fairer dataset [16]. This improves its ability to correctly classify examples of both classes and gives a more accurate picture of how well it works in the underrepresented class, which in this case is kidney disease. Table 1 shows the before and after of the SMOTE implementation of class frequency in the dataset.

Table 1. Class Frequency Comparison before and after Resampling SMOTE.

Resampling using SMOTE	Positive class	Negative class
Before	250	150
After	250	250

2.4 Feature Selection

2.4.1 Correlation Matrix

Before employing the data for training and constructing a predictive model, it is imperative to determine the key factors that exert a substantial influence on the identification of kidney disease events. This crucial process, known as feature selection, is essential for enhancing model accuracy. Among the straightforward techniques for pinpointing influential attributes, correlation analysis takes precedence. This statistical approach quantifies the extent to which two

variables exhibit a linear relationship [17]. For calculating the correlation, Pearson's correlation coefficient stands as the most employed method. The calculation of the Pearson correlation coefficient can be achieved using Equation 2.

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{2}$$

Where, cov(X,Y), $\sigma_X$and σ_Y represent the covariance of X and Y, and the standard deviation for each data X and Y, respectively, as the formulas are shown in Equation 3 and Equation 4 with $\bar{X}$ and $\bar{Y}$ are the average values.

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{X})^2}{N-1}},\ \sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1}} \tag{3}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})}{N-1} \tag{4}$$

The Pearson's correlation coefficient is used to measure the relationship between variables with continuous values. However, for categorical data, some measures of association that can be used to assess the relationship between variables include Tetrachoric, Polychoric, and Cramer's V. In this research, the correlation matrix for categorical variables utilizes Cramer's V as in Equation 5.

$$V = \sqrt{\frac{\chi^2/N}{\min(k-1, r-1)}} \tag{5}$$

2.4.2 Principal Component Analysis

A powerful method for reducing dimensionality, principal component analysis (PCA) has several uses in fields including biology, facial recognition, and classification. The principal component analysis (PCA) method looks for the principal components (PCs) to identify the data's underlying, inherently low-dimensional structure. Due to the fact that the data's dimensionality has been significantly decreased while maintaining its latent structure [18].

The fundamental step of PCA is to transform the correlated mixed signals into a collection of values for the so-called principle components, which are linearly uncorrelated variables. Maximizing the variances of the combined signal is therefore the PCA criteria to estimate the n-th unmixing row vector, producing the n-th main component [19].

$$w_n = \underset{||w||^2=1}{\operatorname{argmax}} J(w^T x_n) \tag{6}$$

$$J(y) = ||y||^2, x_n = (I - \sum_{j=1}^{N-1} w_j w_j^T) \tag{7}$$

x is a created signal from the combined signal x and the unmixing row vector $\{w_j\}_{j=1}^{n-1}$, I is the identity matrix, which has the dimensions $N \times N$, with n = 1, 2,..., N. The first main component has the greatest achievable variance when n = 1. The sequential principal components for n = 2, 3,..., N have the most variance they may with the restriction that they are orthogonal to the preceding principal components, as presented in Equation 6 and Equation 7 [19].

2.5 Prediction Modeling

2.5.1 Logistic Regression

In the healthcare industry, logistic regression is a very well-liked and effective classification technique. The logistic regression model has been used in several published healthcare research on a range of disorders. In the classification of datasets pertaining to kidney disease, logistic linear regression arises as a useful technique. Logistic regression is ideally adapted for binary classification tasks in which the objective is to predict which of two classes an instance belongs to [20]. Using logistic linear regression, the model estimates the probability of an instance belonging to a specific class by combining input features and corresponding weights in a linear fashion. Using a predetermined threshold, this calculated probability is then transformed into a binary outcome [21]. By dividing into two classes, LR of order-n is represented as LR of order-n [22].

$$P(x_1, x_2, \cdots, x_N) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_N x_N)}} \quad (8)$$

In order to estimate the probability values Equation 8, the target characteristic is split into two classes: "successful" and "unsuccessful." It yields a value of 1 for the "successful" class and a value of 0 for the "unsuccessful" class [22]. In the context of classifying kidney diseases, logistic linear regression can learn the relationships between different medical characteristics and whether or not the disease is present. This makes it easier to make accurate predictions based on the characteristics of the dataset. It provides interpretable insights into the impact of individual characteristics on the likelihood of developing kidney disease, making it a valuable diagnostic and risk assessment tool.

2.5.2 Support Vector Machine

In the realm of kidney disease identification, support vector machines (SVM) offer a robust approach to classification. SVM is particularly well-suited for scenarios where data points can be separated by a clear boundary, maximizing the margin between different classes [23]. When applied to kidney disease classification, SVM aims to find a hyperplane that best segregates instances representing the presence or absence of the disease while maximizing the distance between the two classes [24]. The decision boundary is defined by a linear combination of input features, and the model identifies support vectors data points closest to the decision boundary which play a pivotal role in determining the optimal separation with the formula is expressed by Equation 9. The formula used to solve the SVM optimization problem involves minimizing the norm of the weight vector while satisfying certain constraints. Mathematically, this optimization problem can be expressed as in Equation 10 and Equation 11.

$$y(x) = w^t \cdot x + w_0 \quad (9)$$

$$minimize: \min_{w,b} \frac{1}{2} \left|\left|w\right|\right|^2 \quad (10)$$

$$subject\ to: y_i(w \cdot x_i - b) \geq 1\ for\ all\ data\ points\ (x_i, y_i) \quad (11)$$

Here, w represents the weight vector, x_i denotes the input features of the i[th] instance, b is the bias term, and y_i is the corresponding class label (+1 or -1). This formula captures the essence of SVM's goal to find the optimal hyperplane that separates the classes while maintaining the maximum possible margin. In the context of kidney disease identification, SVM's ability to handle complex feature relationships and define clear decision boundaries makes it a valuable asset in accurately classifying instances based on their medical attributes.

2.6 Performance Evaluation

The confusion matrix is used to determine factual information and the classification prediction results of a classification system. The performance of the classification system is generally evaluated using a data matrix. Here is the display of a confusion matrix in Table 2.

Table 2. Confusion Matrix

		Actual	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

The performance evaluation of a model is conducted to assess how well the model performs by using test data for prediction. The evaluation involves calculating accuracy, precision, and sensitivity/recall. The formulas for these performance evaluations [5] are given in Equation 12, Equation 13, and Equation 14.

$$accuracy = \frac{TP + TN}{Total\ Data} \quad (12)$$

$$precision = \frac{TP}{TP + FP} \quad (13)$$

$$f1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (14)$$

2.7 Cross-Validation

A technique for assessing machine learning algorithms is called K-fold cross-validation. The data is divided into a number of K parts using this manner, one of which will be used for testing prior to training. Cross-validation is a method that involves resampling the dataset to evaluate machine learning algorithms with smaller data samples. K-fold cross-validation checks the procedure for each sample after dividing the dataset into K-number of samples, resulting in an average accuracy for the dataset [25].

3. Results and Discussion

The dataset of kidney failure in India from UCI Machine Learning consisting of 400 instances with 25 features were used. The 25 features underwent preprocessing, including transforming categorical features from "yes" and "no" values to 1 and 0, followed by linear interpolation to fill in missing values. Unresolved missing values can lead to changes in the analysis results. Ultimately, data that contains missing values can yield different conclusions compared to cleaned or repaired data.

In the preliminary experiment, SVM and Logistic regression are found to return the highest evaluation scores (accuracy, precision, and F1-score) among other machine learning models for the dataset used in this research. Furthermore, to enhance the recent performance, this study provides three testing scenarios involving hyperparameter model tunning, resampling technique, and feature selection model. Table 3 explains the details of the scenarios provided in this research.

Table 3. Testing Scenarios Description

	The goal	Machine learning model	Applying resampling technique of SMOTE	Missing values treatment	Feature selection
Scenario-1:	to test hyperparameter tuning in improving model performance	SVM and Logistic linear regression.	No	applied	Pearson correlation with threshold.
Scenario-2	to test resampling technique in improving model performance	SVM and Logistic linear regression.	Yes	applied	Pearson correlation with threshold.
Scenario-3	to test PCA feature selection in improving model performance	SVM and Logistic linear regression.	Yes and no	applied	Principal Component Analysis (PCA)

3.1 Testing Scenario

3.1.1 Scenario-1

In Scenario 1, the study attempts to use the SVM and LR methods to test whether hyperparameter tuning improves each model's performance. The feature selection used is performed using Pearson's correlation coefficient, which identifies four significant correlated features. For hyperparameter tuning, GridSearchCV was utilized to obtain the best parameter of each model, SVM and LR.

Table 4. Tuned Parameters

Parameter	Value	
	Support Vector Machine	Logistic Regression
kernel / solver	linear, rbf	liblinear, saga
C	100, 10, 1.0, 0.1, 0.01	100, 10, 1.0, 0.1, 0.01
gamma	scale	none
penalty	none	l1, l2

Table 4 shows the range values of the parameters used for hyperparameter tuning. By using the GridSearchCV method, it obtains the best scores and best parameters. The best score for SVM is 98% from the best parameters of 'C': 10, 'gamma':'scale', and 'kernel': 'linear'. Meanwhile, the LR method resulted in the best score of 98%, with the best parameters being 'C': 10, 'penalty': 'l2', and 'solver': 'liblinear'.

Table 5. The Average and the Deviation of Evaluation Score from K-Fold Validation

Testing	Accuracy		Precision		F1-Score	
	Mean	Deviation	Mean	Deviation	Mean	Deviation
Support Vector Machine	94.25%	2.97%	98.01%	2.63%	95.22%	2.56%
Logistic Regression	94%	2.78%	98.01%	2.63%	95.01%	2.4%
Support Vector Machine Tuning	98%	1.5%	98.49%	2.48%	98.41%	1.17%
Logistic Regression Tuning	98%	2.07%	99.52%	1.43%	98.93%	1.81%

From Table 5, the K-fold cross-validation outcomes for scenario 1 was found, where K equals 10. After applying feature selection using Pearson's correlation coefficient for SVM, the average accuracy, precision, and F1-score are 94.25%, 98.01%, and 95.22% respectively. The corresponding standard deviations are 2.97%, 2.63%, and 2.56%. The execution time for this model is 0.07 seconds as in Table 8. For Logistic Regression, the average accuracy, precision, and F1-score are 94%, 98.01%, and 95.01% respectively, with standard deviations of 2.78%, 2.63%, and 2.4%. The runtime for this model is also 0.07 seconds as in Table 8. Comparing to the results using optimized hyperparameters, which utilize GridSearchCV, the SVM obtains average accuracy, precision, and F1-score are 98%, 98.49%, and 98.41%, respectively. The smallest standard deviation, 1.5%, is observed for accuracy, while precision and F1-score have standard deviations of 2.48% and 1.17% respectively. The runtime for this model is 295 seconds as in Table 8. As for logistic regression, the average accuracy, precision, and F1-score are 98%, 99.52%, and 98.93%, respectively, with standard deviations of 2.07%, 1.43%, and 1.81%. The runtime for this model is 67 seconds as in Table 8. Therefore, hyperparameter tuning significantly increases accuracy, precision, and F1-score on average for both models with a lower deviation of the cross-validation test.

3.1.2 Scenario 2

In Scenario 2, the data were resampled using the SMOTE method to account for imbalances. This was accomplished by oversampling the dataset, which was then used for prediction with SVM and LR. In accordance with Scenario 1, feature selection using the correlation matrix resulted in the identification of four correlated features.

Table 6. The Average and the Deviation of Evaluation Score from K-Fold Validation

Testing	Accuracy		Precision		F1-Score	
	Mean	Deviation	Mean	Deviation	Mean	Deviation
Support Vector Machine	95.2%	2.71%	98.78%	1.87%	94.93%	2.99%
Logistic Regression	94.2%	4.04%	100%	0%	93.64%	4.65%

Table 6 displays the results of k-fold cross-validation for the second scenario, with k maintained at 10. When SMOTE resampling and correlated features were selected, the average accuracy was marginally higher than in

Scenario 1. The average accuracy, precision, and F1-score for the support vector machine are 95.2%, 98.78%, and 94.93%, with standard deviations of 2.71, 1.81, and 2.99%, respectively. This model's execution time remains at 0.04 seconds, reflecting the information presented in Table 8. The average accuracy, precision, and F1-score in the context of logistic regression are 94.2%, 100%, and 93.64%, respectively. The standard deviations are 4.04%, 0%, and 4.56%, respectively. Table 8 depicts that the runtime for this logistic regression model remains slightly quicker at 0.02 seconds.

3.1.3 Scenario 3

In Scenario 3, the study attempts to use the SMOTE method for oversampling the data. Subsequently, feature selection is performed using Principal Component Analysis (PCA) with (n_component = 7) to identify the best features to be predicted by the Support Vector Machine and Logistic Regression methods.

Table 7. The Average and the Deviation of Evaluation Score from K-Fold Validation

Testing	Accuracy		Precision		F1-Score	
	Mean	Deviation	Mean	Deviation	Mean	Deviation
Support Vector Machine	97.5%	2.24%	99.22%	1.57%	97.95%	1.85%
Logistic Regression	98.75%	1.53%	99.52%	1.43%	98.96%	1.28%
Support Vector Machine SMOTE	98.6%	1.28%	99.62%	1.15%	98.57%	1.32%
Logistic Regression SMOTE	98.8%	1.33%	100%	0%	98.77%	1.37%

In Table 7, the results of k-fold cross-validation (k=10) for the third scenario are presented. The first set of results shows the performance without using SMOTE for resampling, but with feature selection using principal component analysis (PCA). It can be observed that there is a considerable improvement in the average and standard deviation for both methods after applying SMOTE with feature selection using PCA. Without SMOTE, the support vector machine achieved an average accuracy, precision, and F1-score of 97.5%, 99.22%, and 97.95% respectively, with standard deviations of 2.24%, 1.57%, and 1.85%. The runtime for this model is 31 seconds as in Table 8. For logistic regression, the average accuracy, precision, and F1-score were 98.75%, 99.52%, and 98.96% respectively, with standard deviations of 1.53%, 1.43%, and 1.28%. The runtime for this model is 0.05 seconds as presented in Table 8.

After applying SMOTE, there was a significant increase in the average accuracy, precision, and F1-score. The support vector machine achieved an average accuracy, precision, and F1-score of 98.6%, 99.62%, and 98.57% respectively, with standard deviations of 1.28%, 1.15%, and 1.32%, which were smaller than the previous results. The runtime for this model is 29 seconds as in Table 8. Similarly, for logistic regression, the average accuracy, precision, and F1-score were 98.8%, 100%, and 98.77% respectively, with standard deviations of 1.33%, 0%, and 1.37%, which were also smaller than the previous results. The runtime for this model is 0.01 seconds as in Table 8. SMOTE led to an increase in the average accuracy, precision, and F1-score, while reducing the standard deviation for both methods.

3.2 Discussion

Three distinct scenarios were conducted, each aimed at evaluating different preprocessing techniques involving the use of linear interpolation for handling missing values. The results exhibited distinct performances for each scenario. Scenario 1 involved hyperparameter tuning, resulting in improved accuracy for both the support vector machine (SVM) and logistic regression methods. Scenario 2 incorporated SMOTE resampling, yielding a minor accuracy boost for both methods. Scenario 3, on the other hand, which combined SMOTE resampling with feature selection using principal component analysis, showed a big improvement in accuracy for both methods, making it the best scenario in this study.

Prior studies in kidney disease prediction through machine learning are noteworthy. For instance, Amirgaliyev et al. utilized support vector machines to forecast chronic kidney disease, achieving 82% accuracy [1]. In this study, the SVM was employed across all three scenarios, and it has outperformed the previous model in terms of average accuracy. Similarly, Ekanayake's study leveraged Little's MCAR method and SVM, attaining 96% accuracy [7]. However, the interpolation-based Scenario 1 produced an average accuracy of 98% for SVM, albeit with a longer

runtime. In Debal's work [9], support vector machines were used alongside feature selection methods, with the unsupervised approach yielding 96.7% accuracy. This study employed principal component analysis for feature selection, achieving an average accuracy of 97.5%, surpassing Debal's model. Likewise, Emon et al. garnered 96% accuracy with principal component analysis and logistic regression [11]. In the Scenario 3, leveraging logistic regression and principal component analysis for feature selection, this study achieved an impressive average accuracy of 98.75%, which further improved to 98.8% with SMOTE resampling. This scenario demonstrated superior accuracy and runtime efficiency.

K-fold cross-validation (K = 10) facilitated various insightful findings across all scenarios. Notably, Scenario 1, featuring SVM with hyperparameter tuning, yielded strong outcomes. Average accuracy, precision, and F1-score were 98%, 98.49%, and 98.41%, with low standard deviations. Although the runtime was relatively lengthy, this contributed to Scenario 3's performance. Table 6 shows that when linear interpolation and principal component analysis were used together in Scenario 3, the results were very accurate and ran quickly. Consequently, Scenario 3 emerges as the most promising among the three.

In Scenario 3, leveraging both SMOTE resampling and principal component analysis for feature selection led to significant enhancements. For SVM, the average accuracy, precision, and F1-score reached 98.6%, 99.62%, and 98.57%, with narrow standard deviations. Logistic regression demonstrated an average accuracy, precision, and F1-score of 98.8%, 100%, and 98.77%, respectively, with equally small standard deviations. This scenario showcased substantial accuracy improvements and reduced standard deviations. Notably, both methods displayed consistent, reliable measurements while maintaining a relatively swift runtime.

Despite its applicability to small datasets, SMOTE's ability to fill data gaps with synthetic samples did not yield optimal results in Scenario 2, where feature selection used the correlation matrix. However, it excelled in Scenario 3 when coupled with principal component analysis, achieving the best outcome. Table 8 explains the details of each runtime model, and scenario 3 is the best model.

Table 8. Model Runtime for Each Scenario

Scenario	Method	Time Consume
Scenario 1	Support Vector Machine	0.07 seconds
	Logistic Regression	0.07 seconds.
	Support Vector Machine Tuning	295 seconds
	Logistic Regression Tuning	67 seconds
Scenario 2	Support Vector Machine	0.04 second
	Logistic Regression	0.02 second.
Scenario 3	Support Vector Machine	31 seconds
	Logistic Regression	0.05 seconds.
	Support Vector Machine SMOTE	29 seconds
	Logistic Regression SMOTE	0.01 seconds

4. Conclusion

This study introduced an automated classification algorithm designed for chronic kidney disease diagnosis based on a comprehensive dataset that encompasses clinical history, physical examination, and laboratory tests. The dataset utilized were sourced from the UCI Machine Learning repository. This study employed SVM and LR prediction models to identify instances of chronic kidney disease. The objective was to assess these models in comparison to the previous related studies. Before entering the prediction models, the dataset underwent missing value treatment using linear interpolation.

Subsequently, a series of test scenarios were conducted, yielding diverse performances. Notably, the third scenario emerged as the most promising. For the SVM, the average accuracy, precision, and F1-score are 98.6%, 99.62%, and 98.57%, respectively, with standard deviations of 1.28%, 1.15%, and 1.32%. Parallelly, logistic regression

produces average accuracy, precision, and F1-score figures of 98.8%, 100%, and 98.77%, with standard deviations of 1.33%, 0%, and 1.37%, respectively. The reliability and consistency of accuracy measurements are shown by the very small standard deviation values for SMOTE resampling and principal component analysis (PCA)-based feature selection. Additionally, when applying principal component analysis after SMOTE resampling, the support vector machine exhibits a runtime of 29 seconds, while logistic regression showcases significantly faster processing at 0.01 seconds.

This study gives a way to improve the speed and accuracy of diagnosing kidney disorders by combining linear interpolation, SMOTE resampling, feature selection, PCA, and prediction models like logistic regression and SVM. This streamlined approach accelerates patient care and diagnostic precision. To further enhance model effectiveness, the expansion of the dataset is recommended. Collecting kidney disease-related data from medical facilities worldwide would bolster the available samples and introduce additional indicators, ultimately enhancing the quality of the kidney disease models generated.

Acknowledgement

The first author acknowledges the support of the Directorate of Research and Community Service at Telkom University through the internal research grant that also partially supports the second author.

References

[1] Y. Amirgaliyev, S. Shamuluulu, and Serek Azamat, “Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods,” IEEE, 2018. https://doi.org/10.1109/ICAICT.2018.8747140

[2] C. K. Leung *et al.*, “Data science for healthcare predictive analytics,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2020. https://doi.org/10.1145/3410566.3410598

[3] G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, *Predictive analysis of diabetic patient data using machine learning and Hadoop*. I-SMAC, 2017. https://doi.org/10.1109/I-SMAC.2017.8058253

[4] K. Deepika and S. Seema, *Predictive analytics to prevent and control chronic diseases*. IEEE, 2016. https://doi.org/10.1109/ICATCCT.2016.7912028

[5] A. Maurya, R. Wable, R. Shinde, S. John, R. Jadhav, and R. Dakshayani, *Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning*. 2019. 10.1109/icnte44896.2019.8946029

[6] F. A. N. Masruriyah, H. H. Handayani, T. Djatna, D. Wahiddin, and K. M. D. Hardhienata, “Predictive Analytics For Stroke Disease,” 2019. https://doi.org/10.1109/ICIC47613.2019.8985716

[7] I. U. Ekanayake and D. Herath, *Chronic Kidney Disease Prediction Using Machine Learning Methods*. 2020. https://doi.org/10.1109/MERCon50084.2020.9185249

[8] S. Pal, “Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms,” *Multimed Tools Appl*, 2023. https://doi.org/10.1007/s11042-023-15188-1

[9] D. A. Debal and T. M. Sitote, “Chronic kidney disease prediction using machine learning techniques,” *J Big Data*, vol. 9, no. 1, Dec. 2022. https://doi.org/10.1186/s40537-022-00657-5

[10] G. M. Ifraz, M. H. Rashid, T. Tazin, S. Bourouis, and M. M. Khan, “Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods,” *Comput Math Methods Med*, vol. 2021, 2021. https://doi.org/10.1155/2021/6141470

[11] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, *Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques*. IEEE, 2016. https://doi.org/10.1109/MITICON.2016.8025242

[12] L. Rubini, P. Soundarapandian, and P. Eswaran, “Chronic_Kidney_Disease,” *UCI Machine Learning Repository*.

[13] G. Huang, “Missing data filling method based on linear interpolation and lightgbm,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. https://doi.org/10.1088/1742-6596/1754/1/012187

[14] D. Thera, S. H. Sitorus, and D. M. Midyanti, “Penerapan Metode Interpolasi Linear dan Histogram Equalization Untuk Perbesaran dan Perbaikan Citra,” *Coding : Jurnal Komputer dan Aplikasi*, vol. 08, 2020.

[15] M. Tahir, F. Khan, M. K. I. Rahmani, and V. T. Hoang, “Discrimination of golgi proteins through efficient exploitation of hybrid feature spaces coupled with smote and ensemble of support vector machine,” *IEEE Access*, vol. 8, pp. 206028–206038, 2020. https://doi.org/10.1109/ACCESS.2020.3037343

[16] A. Ishaq *et al.*, “Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques,” *IEEE Access*, vol. 9, pp. 39707–39716, 2021. https://doi.org/10.1109/ACCESS.2021.3064084

[17] R. Gupta, N. Koli, N. Mahor, and N. Tejashri, *Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease*. International Conference for Emerging Technology (INCET), 2020. https://doi.org/10.1109/INCET49848.2020.9154147

[18] C. M. Feng, Y. Xu, J. X. Liu, Y. L. Gao, and C. H. Zheng, “Supervised Discriminative Sparse PCA for Com-Characteristic Gene Selection and Tumor Classification on Multiview Biological Data,” *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 10, pp. 2926–2937, Oct. 2019. https://doi.org/10.1109/TNNLS.2019.2893190

[19] H. Kwon, W. Q. Malik, S. B. Rutkove, and B. Sanchez, “Separation of Subcutaneous Fat from Muscle in Surface Electrical Impedance Myography Measurements Using Model Component Analysis,” *IEEE Trans Biomed Eng*, vol. 66, no. 2, pp. 354–364, Feb. 2019. https://doi.org/10.1109/TBME.2018.2839977

[20] A. K. Chaudhuri and A. Das, “Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases,” in *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. http://dx.doi.org/10.1109/INOCON50539.2020.9298372

[21] S. H. Adil, M. Ebrahim, K. Raza, and M. A. Hashmani, “Liver Patient Classification using Logistic Regression,” 2018. https://doi.org/10.1109/ICCOINS.2018.8510581

[22] P. Chittora *et al.*, “Prediction of Chronic Kidney Disease - A Machine Learning Perspective,” *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 17312–17334, 2021. https://doi.org/10.1109/ACCESS.2021.3053763

[23] H. Alshamlan, H. Bin Taleb, and A. Al Sahow, “A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression,” in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 38–41. https://doi.org/10.1109/ICICS49469.2020.239549

[24] R. G. Brereton and G. R. Lloyd, "Support Vector Machines for classification and regression," *Analyst*, vol. 135, no. 2. Royal Society of Chemistry, pp. 230–267, 2010. https://doi.org/10.1039/b918972f

[25] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," in *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. https://doi.org/10.1109/ASYU50717.2020.9259880