A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes

Admi Syarif, Olivia Desti Riana, Dewi Asiah Shofiana, Akmal Junaidi
Department of Computer Science-Faculty of Mathematics and Natural Sciences, University of Lampung, Indonesia

***Abstract*—Based on the findings of the 2010 Global Burden of Disease analysis, there was an increase in the global ranking of Chronic Kidney Disease (CKD) as a major contributor to mortality, moving from 27th place in 1990 to 18th position. Approximately 10 percent of the global population experiences CKD, and every year millions of lives are lost due to limited access to adequate treatment. CKD poses a substantial global health concern, greatly affecting both the well-being and life span of individuals afflicted by the condition. This study aims to evaluate the performance of three major classification algorithms in CKD diagnosis: Decision Tree, Support Vector Machine (SVM), and Naïve Bayes. This research distinguishes it from previous studies through an innovative data processing approach. Data preprocessing involved transforming categorical values into numerical form using label encoding, as well as applying Exploratory Data Analysis (EDA) to identify outliers and test data assumptions. In addition, the handling of missing values was done with appropriate strategies to maintain the integrity of the dataset. The classification method was evaluated using a dataset of 400 samples from Kaggle with 24 attributes. Through careful experimentation, the accuracy results of each algorithm are presented and compared. The results of this study can help in the development of a more efficient and accurate decision support system for the early diagnosis of CKD.**

***Keywords*—Chronic kidney disease (CKD); classification; decision tree; machine learning; naïve bayes; support vector machine (SVM)**

I. INTRODUCTION

The kidneys, a pair of bean-shaped organs, are located in the posterior part of the abdomen and play a major role in maintaining the body's internal balance. Their duties include filtering and purifying the blood, eliminating excess fluid and metabolic waste through the formation of urine, as well as regulating electrolyte balance, blood pressure, and the production of hormones that influence the formation of red blood cells. The central role of the kidneys in maintaining body harmony also supports the optimal performance of other organs [1].

Currently, the prevalence of CKD continues to increase globally and has become a serious health problem. Based on the Global Burden of Disease study in 2010, CKD rose to 18th as the world's leading cause of death, up from 27th in 1990. More than two million individuals worldwide undergo dialysis therapy or kidney transplantation, although this number represents only about 10% of the population requiring such treatment. About ten percent of the global population suffers from CKD, and millions of lives are lost each year due to limited access to adequate treatment [2]. Chronic Kidney Disease (CKD) refers to the decline in kidney function that occurs slowly over months or even years [3]. Decreased kidney function can result in the accumulation of fluids, electrolytes, and metabolic waste in the body, which in turn causes various health problems.

In the early stages, CKD often does not cause noticeable symptoms, but patients may experience kidney pain when the disease is in an advanced stage [4]. Chronic kidney failure is progressive and cannot be cured, resulting in a high mortality rate. One of the problems faced by patients with CKD is the high cost of treatment and medication. Therefore, early detection is crucial to identify kidney disease at an early stage and prevent the development of chronic kidney disorders [5].

In the present era, the use of machine learning has become popular in the field of healthcare due to the demand for efficient analytical methodologies to uncover important yet undiscovered information in health data [6]. Medical data mining is employed to gain insights by reviewing information obtained from medical reports, evidence tables, flowcharts, research papers, and more. This data is then transformed into relevant information to support decision-making [7]. Machine learning is a field that encompasses the creation of statistical models and algorithms, empowering computer systems to execute tasks without direct commands, instead of relying on patterns and deduction. By using machine learning algorithms, computer systems can process large amounts of historical data and recognize patterns within that data. This allows the system to make more accurate predictions based on input data.

In this research, three machine learning classification methods are employed, specifically Decision Tree, Support Vector Machine, and Naïve Bayes. The difference from previous studies lies in the preprocessing stage, where several processing techniques are applied to the dataset. One of them is data transformation, where invalid values in categorical data are replaced and categorical values are converted to integers using label encoding. Furthermore, Exploratory Data Analysis (EDA) is conducted, employing descriptive statistics and visual tools to gain a deeper understanding of the data. The

goal of EDA is to uncover maximum insights from the dataset, identify outliers and anomalies, and test underlying assumptions [8]. The missing values are handled by filling in the mean for numerical attributes and the mode for categorical attributes. Additionally, k-fold cross-validation is employed to reduce the impact of accuracy instability. The accuracy is obtained from the average accuracy of each fold [9].

The objective of this research is to implement and evaluate Decision Tree, Support Vector Machine, and Naïve Bayes algorithms in the process of diagnosing CKD. The algorithm is implemented using the Python programming language. The study utilizes a dataset of 400 samples obtained from Kaggle, consisting of 24 attributes. The results of this research will be compared with previous studies to compare different classification methods and conclude the most effective classification.

The article is divided into several sections. In Section I, it covers the background, motivation, related work, and the overall structure of the article. This section provides an overview of the research context, the reasons behind conducting the study, and a review of relevant literature. Section II presents the related works, including the literature review. Section III present the workflow, CKD dataset. It discusses various steps such as preprocessing, handling missing values, Exploratory Data Analysis (EDA), *k*-fold cross-validation, and the implementation of the decision tree, support vector machine, and Naïve Bayes algorithms. This section provides a comprehensive understanding of the dataset and the methodologies employed in the research. The experimental results of the decision tree, SVM, and Naïve Bayes methods in classifying CKD are presented in Section IV. This section evaluates the performance of each classification method and provides insights into their effectiveness in diagnosing the disease. Finally, Section V concludes the paper by summarizing the main findings, highlighting the research contributions, and offering recommendations for future studies.

II. Related Works

Several studies have conducted the classification of CKD using machine learning methods. These are presented in Table I. Based on the information above, previous studies conducted data splitting using the split data function to divide the data into direct training and testing subsets. In this study, we employed the k-fold cross-validation method for data division, which can reduce instability in accuracy. The accuracy is calculated by averaging the accuracy of each fold. Furthermore, this study differed from previous research in terms of data pre-processing, as we applied several data processing techniques to the dataset.

TABLE I. Previous Approaches to Chronic Kidney Disease

Id	Authors	Data	Method	Results (Accuracy)
1	Senan et al. [10]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: University of California	SVM, K-NN, Random Forest.	SVM: 96.67% K-NN: 98.33% **Random Forest: 100%**
2	Saringat et al. [11]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 25 Source: UCI Machine Learning Repository website	SVM, Decision Tree, K-NN, Regression.	SVM: 90.25% Decision Tree: 95.50% K-NN: 94.75% **Regression: 98.25%**
3	Gokiladevi et al. [12]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: UCI Benchmark CKD	K-NN, SVM, Random Forest, Decision Tree, Logistic Regression.	K-NN: 67.50% SVM: 73.75% **Random Forest: 98.75%** Decision Tree: 96.25% Logistic Regression: 94.68%
4	Kumar et al. [13]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: University of California	Decision Tree, Naive Bayes, K-NN, Random Forest, SVM.	Decision Tree: 94.00% Naive Bayes: 93.00% K-NN: 67.00% **Random Forest: 97.00%** **SVM: 97.00%**
5	Tekale et al. [14]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 25 Source: UCI Repository	Decision Tree, SVM.	Decision Tree: 92.00% **SVM: 97.00%**
6	Zeynu [15]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: UCI Repository	K-NN, ANN, Naïve Bayes, Ensamble model.	K-NN: 98.5% ANN: 97.75% Naïve Bayes: 94.5% **Ensamble model: 99.00%**
7	Faddillah et al. [16]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: Indians Chronic Kidney Disease	Naive Bayes.	Naive Bayes: 91.25%
8	Amalia [17]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: UCI Repository	SVM, NN.	**SVM: 95.16%** NN: 93.36%

9	Ariani & Samsuryadi [18]	Data Name: Chronic Kidney Disease Data Count: 400 patients Attributes: 24 Source: UCI Repository Machine Learning Benchmark	K-NN	K-NN: 85.83%

III. METHODS

This research involves several stages in analyzing the performance of CKD classification methods. The process begins with a literature review. The second stage involves collecting the CKD dataset. The third stage is the preprocessing stage, where data transformation, missing value handling, and Exploratory Data Analysis (EDA) are performed. Next is the data partitioning stage, using *k*-fold cross-validation. The subsequent stage involves building the model and finally evaluating the model using a confusion matrix and comparing the results. The stages of analyzing the CKD classification process in this research are illustrated in Fig. 1 below:

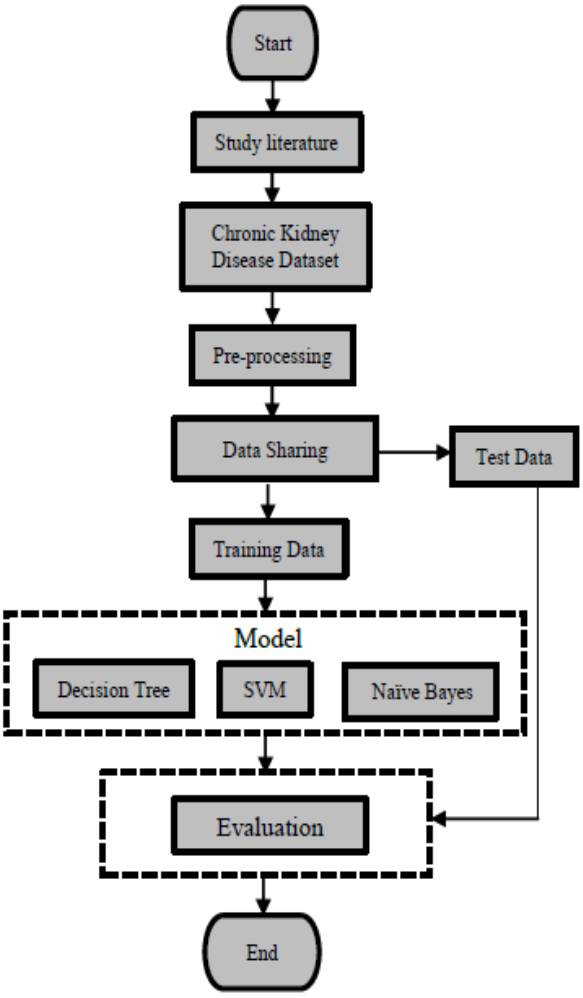

Fig. 1. The research workflow.

A. Dataset

The study utilized data obtained from Kaggle, originating from a Kaggle account created by Nitesh Yadav, a Data Science Intern Technologies from India. The dataset consists of 400 instances with 24 attributes, presented in CSV format. Table II thoroughly describes the attributes related to CKD, providing a comprehensive understanding of the variables involved. By utilizing this dataset, the study aims to analyze and uncover insights regarding the relationships between these attributes and the occurrence of the disease.

TABLE II. DATA FOR CHRONIC KIDNEY DISEASE

Attribute	Possible Values	Types
Age	Years	Numeric
Blood Pressure	mm/Hg	Numeric
Specific Gravity	1.005,1.010,1.015,1.020,1.025	Nominal
Albumin	0,1,2,3,4,5	Nominal
Sugar	0,1,2,3,4,5	Nominal
Red Blood Cells	normal, abnormal	Nominal
Pus cells	normal, abnormal	Nominal
Pus Cell Clumps	present, not present	Nominal
Bacteria	present, not present	Nominal
Blood Glucose Random	mg/dl	Numeric
Blood Urea	mg/dl	Numeric
Serum Creatinine	mg/dl	Numeric
Sodium	mEq/L	Numeric
Potassium	mEq/L	Numeric
Hemoglobin	gms	Numeric
Packed Cell Volume	-	Numeric
White Blood Cell Count	cells/cumm	Numeric
Red Blood Cell Count	millions/cm	Numeric
Hypertension	yes, no	Nominal
Diabetes Mellitus	yes, no	Nominal
Coronary Artery Disease	yes, no	Nominal
Appetite	good, poor	Nominal
Pedal Edema	yes, no	Nominal
Anemia	yes, no	Nominal
Classification	ckd, not ckd	Nominal

B. Pre-processing

1) Data transformation: Incorrect or misleading analysis can occur if there are duplicates or missing data. Therefore, the pre-processing stage plays a crucial role in preparing high-quality data, resulting in more accurate and reliable decisions [19]. In this study, several preprocessing steps were performed on the dataset, including data transformation. During this stage, invalid values in the diabetes mellitus, coronary artery disease, and class attributes were modified. Detailed changes

are recorded in Table III, providing a clear overview of the transformations applied. Through the execution of these data transformations, the dataset gains enhanced accuracy and become well-prepared for subsequent phases of analysis and modeling. Furthermore, to translate categorical attributes into a numerical format, label encoding is employed. In this context, since all categorical attributes have two categories, label encoding can be employed to convert these categorical values into integers, namely 0 and 1.

TABLE III. DATA TRANSFORMATION

Attribute	Transformation
Diabetes mellitus	\tno = no; \tyes = yes; yes=yes
Coronary artery disease	\tno = no
class	ckd\t = ckd; notckd = not ckd

2) Missing value handling: Machine learning models can encounter errors if the dataset contains missing data that is not handled properly. When the dataset is small, discarding samples with missing data is not an appropriate option, as it can reduce the amount of data used to train the machine learning model and affect the accuracy of data analysis. To address this issue, a technique called "Missing Value Handling" is used to handle missing data by filling in appropriate values based on the characteristics of the samples. By filling in the missing data, the dataset can be used to train the machine learning model, resulting in a well-trained model with optimal performance [20]. In this study, the missing data is filled in using the mean and mode values. The mean value is used to fill in missing data in numerical attributes, while the mode value is used to fill in missing data in categorical attributes.

3) Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) is a critical procedure encompassing the recognition and description of repetitive patterns, noteworthy correlation arrangements, and the discernment of variables responsible for noteworthy diversity within a reduced dimensional framework. Moreover, EDA aids in the detection of anomalies like outliers, which might point to potential problems with data quality. It plays a crucial role in understanding the data by uncovering hidden patterns, exploring relationships between variables, and identifying redundant features. EDA serves as an important step in data exploration, enabling researchers to gain insights, make informed decisions, and provide a solid foundation for subsequent analysis and modeling [21].

C. Data Sharing

Within this study, the process of data division is executed through *k*-fold cross-validation. This technique encompasses a series of validation trials where training, validation, and testing phases are carried out. In the initial trial, 80% of datasets chosen at random were employed for training, leaving the remaining 20% for testing purposes. In the subsequent trial, a wholly distinct set of datasets, comprising 80% of the total, is employed for training, while the residual 20% serves for testing purposes. This process is repeated with different sets of 80% training datasets and 20% testing datasets, as shown in Fig. 2, where a total of five experiments are conducted sequentially. Assuming that the selection of training and testing datasets is truly random and the *k*-fold cross-validation process is ergodic, the correct output is obtained by averaging the outputs of all the experiments [9].

Experiment 1	Test	Train	Train	Train	Train
Experiment 2	Train	Test	Train	Train	Train
Experiment 3	Train	Train	Test	Train	Train
Experiment 4	Train	Train	Train	Test	Train
Experiment 5	Train	Train	Train	Train	Test

Fig. 2. K-Fold cross validation.

D. Decision Tree

A conventional tree is composed of a root, branches, and leaves. Similarly, the structure of a Decision Tree includes a root node, branches, and leaf nodes. At each internal node, an attribute is subjected to testing, and the test outcome guides the branch selection, ultimately leading to the assignment of a class label to the corresponding leaf node. Positioned at the highest level, the root node functions as the progenitor of all nodes within the tree. A Decision Tree presents a hierarchical representation where each node signifies a feature (attribute), each link (branch) embodies a decision (rule), and each leaf encapsulates an outcome (categorical or continuous value). Because Decision Trees mimic human cognitive processes, they provide an intuitive means of grasping data and deriving insightful interpretations. The overarching idea is to construct such a tree for the complete dataset, yielding a distinct outcome at each leaf [22].

E. Support Vector Machine

SVM operates as a learning mechanism featuring a hypothesis space founded on linear functions within a feature space of significant dimensions. Its training is facilitated by learning algorithms rooted in optimization theory principles. The accuracy level achieved by the SVM model is highly dependent on the kernel function and parameters used during the training process. Based on its characteristics, the SVM method can be divided into two types: Linear SVM and Non-linear SVM. Linear SVM separates data linearly by placing a hyperplane with a soft margin between classes. The illustration of the linear SVM can be seen in Fig. 3. On the other hand, Non-linear SVM implements the kernel trick by mapping the data into a higher-dimensional space [23]. Basically, the concept of SVM involves finding the optimal separator on the hyperplane, as shown in Fig. 4. The best-separating hyperplane is determined by searching for the value of $f(x)$ on the hyperplane margin [24-26].

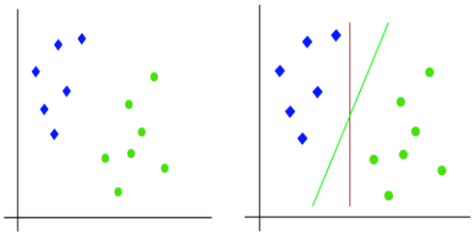

Fig. 3. Linear SVM [25].

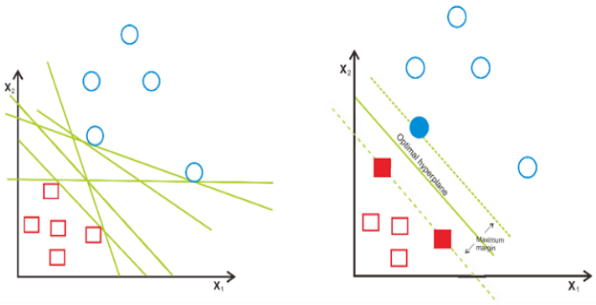

Fig. 4. The effort to find the best hyperplane [26].

F. Naïve Bayes

Naive Bayes classification employs the principle of maximum likelihood estimation to categorize samples into the most probable groups [27]. Given an input vector with features represented as *X* and a class label denoted as *Y*, the concept of Naive Bayes is symbolized by *P(Y|X)*, indicating the probability of class label Y considering the observed features *X*. This representation signifies the posterior probability of *Y*. The original probability *P(Y)*, recognized as the prior probability, is also considered. Throughout the training phase, the task involves acquiring knowledge about the posterior probabilities (*P(Y|X)*) for every combination of *X* and *Y* using insights gleaned from the training dataset [28].

G. Evaluation Metrics

The assessment conducted in this study employs a Confusion Matrix. The Confusion Matrix serves as a performance evaluation tool for machine learning classification tasks encompassing two or more classes. This table showcases various combinations of projected and actual values. Comprising four terms, the Confusion Matrix outlines the classification outcomes: True Positive, True Negative, False Positive, and False Negative [29]. True Positive (TP) signifies accurate positive predictions, while True Negative (TN) signifies accurate negative predictions. False Positive (FP) relates to an erroneous positive prediction, whereas False Negative (FN) corresponds to an incorrect negative prediction. The Confusion Matrix entails several computations, including:

Accuracy measures how accurately a model classifies data correctly [29]. The calculation of accuracy is done using by using Eq. (1).

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \times 100\% \quad (1)$$

Precision describes the accuracy between the requested data and the predicted results given by the model [29]. The calculation of Precision is done using Eq. (2).

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP}+\text{FP})} \times 100\% \quad (2)$$

Recall or Sensitivity represents the success of the model in capturing information [29]. The calculation of Recall or Sensitivity is done using Eq. (3).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100\% \quad (3)$$

F-1 Score represents the weighted average of Precision and Recall. Accuracy is used as a performance reference for algorithms when the dataset has a significant number of false negatives and false positives. However, if the numbers are not close, the *F*-1 Score is used as a study in [29]. The calculation of the F-1 Score is done using Eq. (4).

$$F-1\ \text{Score} = 2x\frac{(\text{Recall x Precision})}{(\text{Recall}+\text{Precision})} \times 100\% \quad (4)$$

IV. Experiments and Results

In this study, we utilized the laptop on the Windows 10 64-bit operating system with 8 GB of memory. It is powered by an Intel(R) Core (TM) i5-4210U processor, which provides processing capabilities ranging from 1.70 GHz to 2.39 GHz. The software employed for the study includes Python 3.10.6, Anaconda Navigator, Jupiter Notebook, a web browser, and Microsoft Excel. These tools and technologies formed the computational environment in which the analyses and experiments were conducted.

For the experiments, the data is divided by using used 5-Fold cross validation (80% data for training and 20% data for testing). The following Table IV presents the comparison of the method for each fold.

TABLE IV. The Value of K-Fold Cross-Validation in a Model

Fold	**Decision Tree**	**Support Vector Machine**	**Naïve Bayes**
	Accuracy		
1	0.975	0.9625	1.00
2	0.9875	0.975	0.975
3	0.95	0.95	0.925
4	0.975	1.0	0.95
5	0.9875	1.0	0.9375
Average	**0.975**	**0.9775**	**0.9575**

TABLE V. Confusion Matrix Decision Tree

Class	Positive predicted	Negative predicted
Positive actual	245	5
Negative actual	5	145

In the above Table V, we show the confusion matrix to decision tree. There are 245 instances of true positives (TP), representing cases that truly belong to the positive class of CKD and are accurately identified as such by the CKD prediction. There are five false positives (FP), which are samples that actually belong to the positive class of having

CKD but are incorrectly predicted as negative for CKD. There are five false negatives (FN), which are samples that actually belong to the negative class but are incorrectly predicted as positive. There are 145 true negatives (TN),

TABLE VI. CONFUSION MATRIX FOR SVM

Class	Positive predicted	Negative predicted
Positive actual	246	4
Negative actual	5	145

Table VI presents the confusion matrix for support vector machine. There are 246 true positives (TP), which are samples that actually belong to the positive class of having CKD and are correctly predicted as positive for CKD. There are four false positives (FP), which are samples that actually belong to the positive class of having CKD but are incorrectly predicted as negative for CKD. There are five false negatives (FN), which are samples that actually belong to the negative class but are incorrectly predicted as positive for CKD. There are 145 true negatives (TN).

The above Table VII shows the confusion matrix for naïve Bayes, there are 233 true positives (TP. There are 17 false positives (FP), which are samples that actually belong to the positive class of having CKD but are incorrectly predicted as negative for CKD. There are 0 false negatives (FN), which means there are no samples that actually belong to the negative class of not having CKD but are incorrectly predicted as positive for CKD. There are 150 true negatives (TN), which are samples that actually belong to the negative class of not having CKD and are correctly predicted as negative for CKD.

Table VIII presents the comparison of these three models. It is shown that the SVM method performs exceptionally well, exhibiting excellent precision, recall, F1-score, and accuracy, making it highly reliable for diagnosing CKD. The Naïve Bayes method yields good results, although slightly lower than SVM and Decision Tree in some metrics, with high precision, but a slightly lower recall. However, the Naïve Bayes method still demonstrates good overall performance with a relatively high F1-score and accuracy in predicting CKD. All three methods show satisfactory performance in diagnosing CKD, with the SVM method standing out as the most accurate.

Finally, here, we also compared our results with those of previous studies. We summarize the comparison in Fig. 5.

TABLE VII. CONFUSION MATRIX NAÏVE BAYES

Class	Positive predicted	Negative predicted
Positive actual	233	17
Negative actual	0	150

TABLE VIII. THE COMPARISON OF THE METHODS

Method	Precision	Recall	F1-score	Accuracy
Decision Tree	0.97	0.97	0.97	0.97
SVM	0.98	0.98	0.98	0.98
Naïve Bayes	0.96	0.96	0.96	0.96

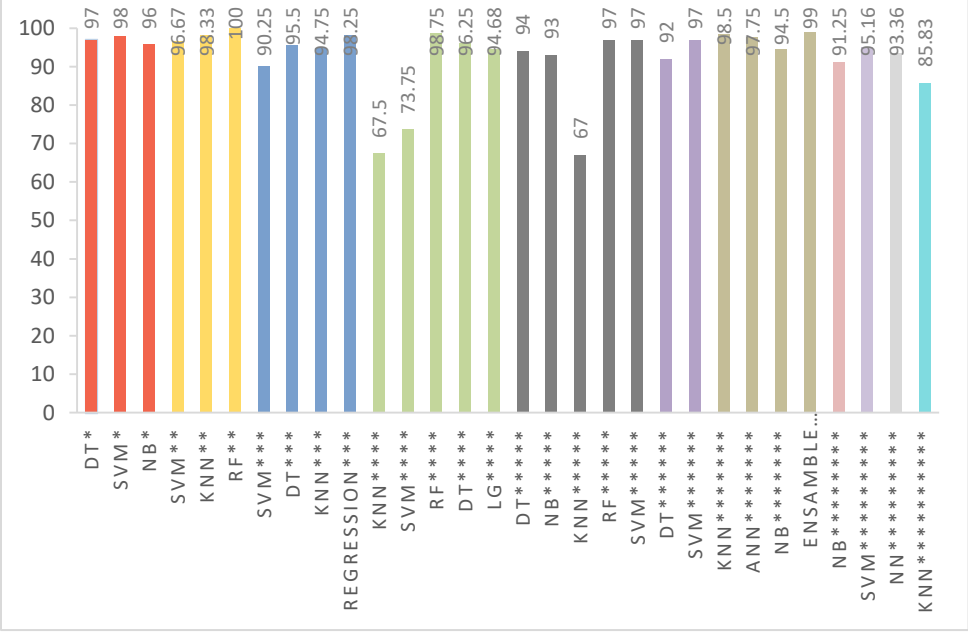

Fig. 5. Comparison graph with several previous studies.

Description :

*/ This research
**/ [10]
***/ [11]
****/ [12]
*****/ [13]
******/ [14]
*******/ [15]
********/ [16]
*********/ [17]
**********/ [18]

V. Conclusion and Future Works

A. Conclusion

Through meticulous implementation, we successfully applied three major classification algorithms in the CKD diagnosis process, namely Decision Tree, Support Vector Machine (SVM), and Naïve Bayes. The analysis of the results confirms that all three methods have exhibited exceptional performance in predicting the disease. Notably, the most intriguing outcome is the achievement of the highest accuracy rate by the Support Vector Machine (SVM) method, reaching a score of 0.98. Thus, this research not only provides deeper insights into early diagnosis and CKD management but also offers valuable guidance in utilizing the most effective classification algorithms for this condition.

B. Recommendations

The recommendations provided in this study are as follows:

1) For future research, it is suggested to explore newer classification methods such as deep learning or ensemble learning. The use of more complex classification methods, may provide advantages in diagnosing CKD.

2) Future studies could utilize larger datasets encompassing a wider attribute range. This will aid in better analyzing the performance of classification methods. Additionally, considering data from other sources to supplement the analysis could be beneficial.

Acknowledgment

A "Fundamental research Grant" from Indonesian Government supported this research. The dataset was obtained from the Kaggle platform, and we acknowledge Nitesh Yadav for providing us with this valuable dataset for this study.

References

[1] H. Kusuma and Suhartini, "Understanding Chronic Kidney Disease and its Treatment." Semarang: Faculty of Medicine, Diponegoro University, 2019.

[2] Aulia, "Chronic Kidney Disease," p2ptm.kemkes.go.id, 2017.

[3] Ministry of Health Indonesia (Kemenkes RI), "What Are the Functions of the Kidneys?," p2ptm.kemkes.go.id, 2019.

[4] Ministry of Health Indonesia (Kemenkes RI), "Symptoms of Chronic Kidney Disease Often Go Unnoticed, Suddenly Stage 5," www.kemkes.go.id, 2023.

[5] Sehat Negriku Sehat Bangsaku, "Maintaining Healthy Kidneys," sehatnegeriku.kemkes.go.id, 2016.

[6] S. F. Sung, P. J. Lee, C. Y. Hsieh, and W. L. Zheng, "Medication use and the risk of newly diagnosed diabetes in patients with epilepsy: A data mining application on a healthcare database," J. Organ. End User Comput., vol. 32, no. 2, pp. 93–108, 2020, doi: 10.4018/JOEUC.2020040105.

[7] S. Kiruthika Devi, S. Krishnapriya, and D. Kalita, "Prediction of heart disease using data mining techniques," Indian J. Sci. Technol., vol. 9, no. 39, 2016, doi: 10.17485/ijst/2016/v9i39/102078.

[8] M. Batta, "Machine Learning Algorithms - A Review," Int. J. Sci. Res., vol. 18, no. 8, pp. 381–386, 2018, doi: 10.21275/ART20203995.

[9] F. Y. H. Ahmed, Y. H. Ali, and S. M. Shamsuddin, "Using K-fold cross validation proposed models for SpikeProp learning enhancements," Int. J. Eng. Technol., vol. 7, no. 4, pp. 145–151, 2018, doi: 10.14419/ijet.v7i4.11.20790.

[10] E. M. Senan, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/1004767.

[11] Z. Saringat, A. Mustapha, R. D. R. Saedudin, and N. A. Samsudin, "Comparative analysis of classification algorithms for chronic kidney disease diagnosis," Bull. Electr. Eng. Informatics, vol. 8, no. 4, pp. 1496–1501, 2019, doi: 10.11591/eei.v8i4.1621.

[12] M. Gokiladevi, S. Santhoshkumar, and V. Varadarajan, "Machine Learning Algorithm Selection for Chronic Kidney Disease Diagnosis and Classification," Malaysian J. Comput. Sci., vol. 2022, no. Special Issue 1, pp. 102–115, 2022, doi: 10.22452/mjcs.sp2022no1.8.

[13] S. Kumar and Mohammad, "Intelligent Systems with Applications Chi 2 -MI: A hybrid feature selection based machine learning approach in the diagnosis of chronic kidney disease," Intell. Syst. with Appl., vol. 16, no. October, p. 200144, 2022, doi: 10.1016/j.iswa.2022.200144.

[14] S. Tekale, P. Shingavi, S. Wandhekar, and A. Chatorikar, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm," no. October 2018, 2022, doi: 10.17148/IJARCCE.2018.71021.

[15] S. Zeynu, "Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method," vol. 15, pp. 168–176, 2018.

[16] A. N. Faddillah, J. Wijaya, and R. Hidayat, "Application of Naive Bayes Algorithm for the Diagnosis of Chronic Kidney Disease," J. Inf., vol. 18, no. 2, pp. 102–106, 2019, doi: 10.36054/jict-ikmi.v18i2.69.

[17] H. Amalia, "Comparison of Data Mining Methods SVM and NN for the Classification of Chronic Kidney Disease," vol. 14, no. 1, pp. 1–6, 2018.

[18] A. Ariani and Samsuryadi, "Classification of Chronic Kidney Disease Using K-Nearest Neighbor," Annual Research Seminar Proceedings, vol. 5, no. 1, pp. 148–151, 2019.

[19] H. Yang, "Data Preprocessing-Chapter 3," Citeseerx, 2013.

[20] R. A. Maula and Gunawan, "Handling Missing Value with Regression Approach in Small-Scale Aquaculture Dataset," J. Rekayasa Elektr., vol. 18, no. 3, pp. 175–184, 2022, doi: 10.17529/jre.v18i3.25903.

[21] V. Da Poian and B. Theiling, "Exploratory Data Analysis (EDA) Machine Learning Approaches for Ocean World Analog Mass Spectrometry," Front. Astron. Sp. Sci., vol. 10, no. May, pp. 1–17, 2023, doi: 10.3389/fspas.2023.1134141.

[22] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," Int. J. Comput. Sci. Eng., vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[23] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Classification of Oral and Dental Diseases Using Support Vector Machine Method," J-Ptiik, vol. 2, no. 2, pp. 802–810, 2018.

[24] E. A. Kurnianto, I. Cholissodin, and E. Santoso, "Classification of Chronic Kidney Disease Patients Using Support Vector Machine (SVM) Algorithm," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 12, pp. 6597–6602, 2018.

[25] Trivusi, "Complete Explanation of Support Vector Machine (SVM) Algorithm," trivusi.web.id, 2022.

[26] A. P. Wibawa, M. G. A. Akbar, P. M. Fathony, and F. A. Dwiyanto, "metode metode klasifikasi.pdf," Proceedings of the Seminar on Computer Science and Information Technology, vol. 5, 2018.

[27] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved Naive Bayes Classification Algorithm for Traffic Risk Management," EURASIP J. Adv. Signal Process., vol. 2021, no. 1, 2021, doi: 10.1186/s13634-021-00742-6.

[28] E. Prasetyo, R. A. D. Rahajoe, and A. Arizal, "Comparison of K-Support Vector Nearest Neighbour and Decision Tree to Naive Bayes," National Seminar on Information Technology Proceedings, pp. 1–6, 2013.

[29] M. S. Anggreany, "Confusion Matrix," socs.binus.ac.id, 2022.