Paradigma, Vol. 26, No. 1, Maret 2024, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 58-64
Published: LPPM Universitas Bina Sarana Informatika
PARADIGMA
website: http://jurnal.bsi.ac.id/index.php/paradigma/

Building a Predictive Model for Chronic Kidney Disease: Integrating KNN and PSO

Slamet Widodo[1], Herlambang Brawijaya[2], Samudi[3]

[1,2] Universitas Bina Sarana Informatika, Jakarta, Indonesia
[3] Universitas Nusa Mandiri, Jakarta, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: February 28, 2024
Revised: March 18, 2024
Accepted: March 29, 2024

Keyword:

K-Nearest Neighbors
Predictive Model
Chronic Kidney Disease
Particle Swarm Optimization

ABSTRACT

This study examines the improvement of prediction accuracy for Chronic Kidney Disease (CKD) through the integration of the K-Nearest Neighbors (KNN) method with Particle Swarm Optimization (PSO). Amidst the rising prevalence of CKD, closely related to diabetes and hypertension, early detection of CKD becomes a significant challenge, especially in Indonesia where access to healthcare facilities and public awareness remain limited. This study utilizes the Chronic Kidney Disease dataset from the UCI Machine Learning repository, encompassing 400 patient records with 24 clinical, laboratory, and demographic variables. With the KNN method, this approach classifies data based on feature proximity, while PSO is used for feature selection and parameter optimization, enhancing the model's accuracy and efficiency in identifying CKD at early stages. The findings indicate a significant improvement in prediction accuracy, from 80.00% using KNN to 97.75% after integration with PSO. These results affirm that the combined approach of KNN and PSO holds great potential in improving early detection and management of CKD, paving the way for further research into practical applications in the healthcare field.

Corresponding Author:

Slamet Widodo,
Sistem Informasi,
Universitas Bina Sarana Informatika,
Jln. Kramat Raya No. 98, Jakarta, (021) 8000063,
Email: slamet.smd@bsi.ac.id.

INTRODUCTION

Chronic kidney disease (CKD) represents a gradual decline in kidney function over time and is chronic in nature. Globally, it's estimated that 8-16% of the population is affected by this disease, with diabetes and hypertension being the primary causes. In Indonesia, the prevalence of CKD is also on the rise, in line with the increasing cases of diabetes and hypertension. Several challenges in managing CKD in Indonesia include limited access to adequate healthcare facilities, public awareness about kidney disease, and the high cost of treatment. This disease often goes undetected until it reaches advanced stages. Early detection through blood and urine tests can assist in managing this condition. Research and education efforts on CKD are also crucial in reducing the burden of this disease in Indonesia. Various studies on CKD have been conducted using different methods and techniques.

(Rady & Anwar, 2019) Evaluated four different models with Probabilistic Neural Networks achieving the highest accuracy (96.70%), followed by Support Vector Machine (87.00%). Multilayer Perceptron and Radial Basis Function showed lower performance, with accuracies of 60.70% and 51.50%, respectively. (Arif-Ul-Islam & Ripon, 2019) Explored the use of boosting algorithms with LogitBoost reaching the highest accuracy (99.75%). AdaBoost, J48, and Ant-Miner also produced excellent results, while SVM and NB showed lower performance. (Alaiad, Najadat, Mohsen, & Balhaf, 2020) tested various algorithms with K-nearest neighbors achieving the highest accuracy (98.50%). Other algorithms like Support Vector Machine, Naïve Bayes, Decision Tree, and JRip also showed competitive accuracy.

DOI: https://doi.org/10.31294/p.v26i1.3282

(Pramanik, Khare, & Gourisaria, 2021) Found results for Random Forest with an accuracy of 98.81%. (Rezayi, Maghooli, & Saeedi, 2021) tested various algorithms, where Random Forest achieved the highest accuracy (99.09%). Deep Learning and Neural Network also showed impressive results with accuracies of 98.04% and 96.52%, respectively. Other algorithms like Naive Bayes, Support Vector Machine, KNN, DT, and Multilayer Perceptron showed accuracy variations from 71.56% to 96.22%. (Saha, Gourisaria, & Harshvardhan, 2022) Showed that almost all tested models had accuracies below 99.08%, with Random Forest reaching exactly 99.08%. (Purwaningsih, 2022) focused on SVM with Feature Selection (FS), finding that the combination of SVM (radial)+FS yielded the best results with an accuracy of 99.75%, followed by SVM (dot)+FS (99.50%) and SVM (polynomial)+FS (95.50%). (Wijaya, 2024) Focused on SVM with PSO, slightly different from FS, and provided the best results with an accuracy of 99.75%.

The conclusion from the conducted research indicates a significant variation in accuracy among various machine learning algorithms for CKD prediction. Models like K-nearest neighbors, SVM with FS, and various boosting techniques showed high accuracy. These results affirm the importance of selecting the right algorithm and tailoring it to data features to improve the performance of predictive models. The problem of delayed early detection of Chronic Kidney Disease (CKD) due to lack of awareness and varying accuracy resulting in CKD predictions using algorithms that have been carried out previously shows the need to improve prediction methods, especially the KNN method, so that early detection is easier and management of this disease is effective.

The approach or solution proposed in this research suggests the integration of two machine learning techniques, namely K-Nearest Neighbors (KNN) and Particle Swarm Optimization (PSO), to develop a more accurate predictive model for CKD. KNN is known for its simplicity in classifying data based on feature proximity, while PSO will be used to optimize feature selection and parameters on KNN (Ariyati et al., 2020), aiming to enhance accuracy and efficiency of the model in identifying CKD at early stages (Ridwansyah, Riyanto, Hamid, Rahayu, & Purnama, 2022). Innovation in this research may allow for earlier and more accurate detection of CKD, potentially saving more lives and reducing the economic burden of this disease in Indonesia and worldwide.

RESEARCH METHOD

Data mining contributes to identifying patterns and trends in the disease along with its treatment solutions and facilitates healthcare professionals in providing more accurate and efficient care. This allows for earlier preventive measures against more serious conditions. In this research method, there are research steps using the KNN and PSO methods, as shown in Figure 1.

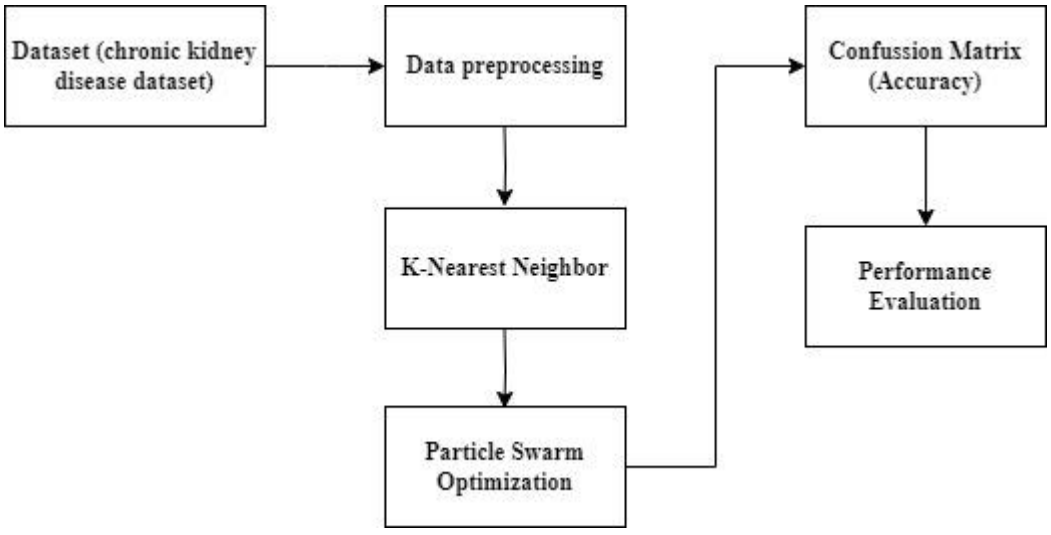

Figure 1. Flow Diagram for predicting CKD

In Figure 1, the process begins with data collection, where data related to chronic kidney disease are gathered for analysis. This initial stage demands the comprehensive and multifaceted acquisition of data on chronic kidney disease, covering clinical, laboratory, and demographic variables of patients. This forms a critical foundation, where the accuracy and completeness of the data are key to ensuring the validity of the predictive model to be developed. This research utilizes secondary data titled "Chronic Kidney Disease," which is taken from the UCI Machine Learning repository and available at https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. The dataset consists of 400 entries with 24 variables or attributes, including one label indicating whether the subject suffers from CKD or not. The CKD

data is included in Table 1.

Table 1. Dataset Information

Attribute	Value
age	Age in years
bp	Blood Pressure (bp in mm/Hg)
Sg	Specific Gravity (1.005,1.010,1.015,1.020,1.025)
Al	Albumin (0,1,2,3,4,5)
Su	Sugar (0,1,2,3,4,5)
Rbc	Red Blood Cells (1 normal,2 abnormal)
Pc	Pus Cell (1 normal,2 abnormal)
Pcc	Pus Cell clumps (1 present,2 notpresent)
Ba	Bacteria (1 present,2 notpresent)
Bgr	Blood Glucose Random (bgr in mgs/dl)
Bu	Blood Urea (bu in mgs/dl)
Sc	Serum Creatinine (sc in mgs/dl)
Sod	Sodium (in mEq/L)
Pot	Potassium (in mEq/L)
Hemo	Hemoglobin (in gms)
Pcv	Packed Cell Volume
Wc	White Blood Cell Count (in cells/cumm)
Rc	Red Blood Cell (in millions/cmm)
Htn	Hypertension (1 yes,2 no)
Dm	Diabetes Mellitus (1 yes,2 no)
Cad	Coronary Artery Disease (1 yes,2 no)
appet	Appetite (1 good,2 poor)
Pe	Pedal Edema (1 yes,2 no)
ane	Anemia (1 yes,2 no)
class	(ckd,notckd)

In Table 1 below, the data undergo Preprocessing Data to clean and prepare it for analysis. This step involves a series of complex data purification procedures including handling missing values, normalization, and data transformation. The purpose of this process is to simplify the initially unstructured data set into a more structured format that is easily interpretable by machine learning algorithms. Subsequently, the data is divided into the data splitting stage (Training & Testing) to prepare the dataset to be used in training and testing the model for classification. For the classification, the K-Nearest Neighbor (KNN) method is applied to the training data to develop a predictive model based on similar features, while Particle Swarm Optimization (PSO) is used for optimization to improve the effectiveness of the model by adjusting parameters automatically (Nurdin, Sartini, Sumarna, Maulana, & Riyanto, 2023). The performance evaluation of the model is conducted using a confusion matrix to measure accuracy (Ridwansyah, Wijaya, & Purnama, 2020). The model integrated with PSO is then evaluated in the Model Evaluation stage to measure the performance of the model. Finally, the results of the evaluation are interpreted in the interpretation stage to understand the implications of the model for chronic kidney disease detection.

K-Nearest Neighbor is a learning technique based on instances, where the training dataset is stored and used to find similarities between new entries and existing ones by comparing their features (Larose & Larose, 2015). In this method, a new entry is classified based on its similarity to one or more nearest neighbors in the dataset. If using more than one neighbor, the classification is determined by the majority or weighted average of the k nearest neighbors (Witten, 2017). The effectiveness of this algorithm is influenced by the quality of data and the selection of the appropriate number of nearest neighbors, making it effective for datasets with noise but requiring careful consideration in parameter selection. KNN is not only used for classification but also for regression (Sarker, 2021).

K-Nearest Neighbor (KNN) enriched with the Particle Swarm Optimization (PSO) algorithm presents a more dynamic and adaptive approach in determining the optimal number of nearest neighbors (k). PSO, an optimization technique inspired by the social behavior of birds or fish, is used to find the optimal values of parameters in a problem, including the selection of k in KNN (Iqbal et al., 2020).

RESULTS AND DISCUSSION

The initial step involves gathering comprehensive data on chronic kidney disease, including clinical, laboratory, and demographic information that is crucial for building an accurate predictive model. This requires data cleansing, such as addressing missing data and normalization as seen in Table 2.

Tabel 2. The CKD Data After Preprocessing

age	69	48	59	56	40	23	45	57	51	34
bp	70	110	70	90	80	80	80	80	60	80
sg	1.010	1.015	1.010	1.010	1.025	1.025	1.025	1.025	1.025	1.025
al	4	3	1	4	0	0	0	0	0	0
su	3	0	3	1	0	0	0	0	0	0
rbc	1	0	0	1	1	1	1	1	1	1
pc	0	1	0	0	1	1	1	1	1	1
pcc	1	1	0	1	0	0	0	0	0	0

ba	1	0	0	0	0	0	0	0	0	0
bgr	214	106	424	176	140	70	82	119	99	121
bu	96	215	55	309	10	36	49	17	38	27
sc	6.3	15.2	1.7	13.3	1.2	1.0	0.6	1.2	0.8	1.2
sod	120	120	138	124	135	150	147	135	135	144
pot	3.9	5.7	4.5	6.5	5.0	4.6	4.4	4.7	3.7	3.9
hemo	9.4	8.6	12.6	3.1	15.0	17.0	15.9	15.4	13.0	13.6
pcv	28	26	37	9	48	52	46	42	49	52
wc	11500	5000	10200	5400	10400	9800	9100	6200	8300	9200
rc	3.3	2.5	4.1	2.1	4.5	5.0	4.7	6.2	5.2	6.3
htn	1	1	1	1	0	0	0	0	0	0
dm	1	0	1	1	0	0	0	0	0	0
cad	1	1	1	0	0	0	0	0	0	0
appet	1	1	1	0	1	1	1	1	1	1
pe	1	0	0	1	0	0	0	0	0	0
ane	1	1	0	1	0	0	0	0	0	0
class	1	1	1	1	0	0	0	0	0	0

After the data preprocessing stage is completed as seen in Table 2, the data is then ready for analysis using the K-Nearest Neighbors (K-NN) model. By employing this method, we can generate a Confusion Matrix, a tool for measuring the effectiveness of classification models including K-NN. The Confusion Matrix provides details about the actual performance of the model compared to its prediction results. The performance of this model can be measured through parameters such as accuracy, the results of which are displayed in Table 3 using the K-NN approach.

Table 3. Performa K-NN

	true ckd	**true notckd**
pred ckd	197	27
pred notckd	53	123

In Table 3, it is known that True CKD represents cases that actually have chronic kidney disease. True notCKD represents cases that actually do not have chronic kidney disease. Pred CKD represents cases predicted by the model as having chronic kidney disease. Pred noCKD represents cases predicted by the model as not having chronic kidney disease. Based on the explanation above, the values in the table can be interpreted as follows:

The count of 197: This is the number of cases that truly have chronic kidney disease and are also predicted by the model to have chronic kidney disease (True Positive - TP). This means the model correctly identified 197 cases of chronic kidney disease.

The count of 27: This is the number of cases that do not have chronic kidney disease but are falsely predicted by the model to have chronic kidney disease (False Positive - FP). This means the model incorrectly identified 27 cases without chronic kidney disease as having chronic kidney disease.

The count of 53: This is the number of cases that actually have CKD but are predicted by the model as not having CKD (False Negative - FN). This means there are 53 cases of chronic kidney disease that the model failed to identify.

The count of 123: This is the number of cases that actually do not have chronic kidney disease and are also predicted by the model to not have chronic kidney disease (True Negative - TN). This means the model correctly identified 123 cases without chronic kidney disease.

After the testing stage with the K-NN model, the next integration stage will be performed. This integration stage combines the strengths of both algorithms: KNN provides the framework for classification, while PSO optimizes the parameters of KNN to fit the unique characteristics of the chronic kidney disease dataset. This integration results in Attribute Weight, and the performance of this model can be measured through parameters such as accuracy, which are displayed in Table 4 and Table 5 using the K-NN and PSO approaches.

Table 4. Attribute Weight Values of KNN+PSO Model

Attribute	**Weight**
age	0.0
Bp	0.0
Sg	0.694

Al	0.0
Su	1.0
rbc	0.0
Pc	0.404
pcc	0.560
Ba	0.0
bgr	0.0
Bu	0.026
Sc	0.0
sod	1.0
pot	1.0
hemo	1.0
pcv	0.0
wc	0.0
rc	0.0
htn	0.0
dm	1.0
cad	1.0
appet	1.0
pe	1.0
ane	1.0

In Table 4, it is observed that attributes Age, Bp, Rbc, Ba, Bgr, Sc, Pcv, Wc, Rc, and Htn with a weight of 0.0 indicate that these attributes do not significantly contribute to predicting chronic kidney disease in this model. This could be due to low variability, redundant information, or lack of significant correlation with the outcome of chronic kidney disease. The attribute Sg with a weight of 0.694 indicates a fairly significant contribution. Sg is an important indicator of kidney function, with lower values indicating potential kidney disease. Although albumin (Al) has a weight of 0.0, blood urea (Bu) with a weight of 0.026 has a small contribution, suggesting that blood urea may slightly influence the model's prediction. High blood urea levels may indicate kidney dysfunction. Pc with a weight of 0.403 shows moderate contribution. The presence of pus cells in urine can be an indicator of kidney infection or inflammation. Pcc with a weight of 0.560 shows a more significant contribution than Pc, indicating the importance of detecting clumps of pus cells in diagnosing kidney disease. Sod, Pot, Hemo, Dm, Cad, Appet, Pe, Ane - all these attributes have a weight of 1.0, indicating a very significant contribution to predicting chronic kidney disease. This reflects the importance of these parameters in assessing kidney condition and overall health, including electrolyte balance, blood condition, and comorbidities.

The weights assigned to each attribute in this predictive model depict the importance of these variables in diagnosing chronic kidney disease. Attributes with higher weights are considered more critical in determining prediction outcomes, while those with lower or zero weights are less significant. The integration of KNN and PSO allows dynamic adjustment of these weights to enhance the accuracy of the predictive model, highlighting how optimization techniques can improve the performance of predictive models in the medical field. From these weights, the model's performance with PSO can be observed.

Table 5. Performa K-NN+PSO

	true ckd	**true notckd**
pred ckd	241	0
pred notckd	9	150

Based on the explanation of Table 5 above, the values in the table can be interpreted as follows:

Count of 241: This is the number of cases that truly have chronic kidney disease and are also predicted by the model to have chronic kidney disease (True Positive - TP). This means the model correctly identified 241 cases of chronic kidney disease.

Count of 9: This is the number of cases that actually have CKD but are predicted by the model as not having CKD (False Negative - FN). This means there are 9 cases of chronic kidney disease that the model failed to identify.

Count of 150: This is the number of cases that actually do not have chronic kidney disease and are also predicted by the model to not have chronic kidney disease (True Negative - TN). This means the model correctly identified 150 cases without chronic kidney disease.

After the model optimization is completed, evaluation will be conducted through testing and validation processes on the results obtained from both methods. Both the K-NN and K-NN+PSO algorithms yield comparison values that can be observed in Table 6.

Table 6. The Accuracy Results of Method Comparison

Method	Accuracy
K-NN	80,00
K-NN+PSO	97,75

Based on the comparison in Table 6, it is evident that

the K-NN method experiences an increase in accuracy after integration using the PSO method, from an accuracy value of 80.00% to 97.75%. This increase in accuracy can also be observed in Figure 2, which illustrates the comparison graph of these values

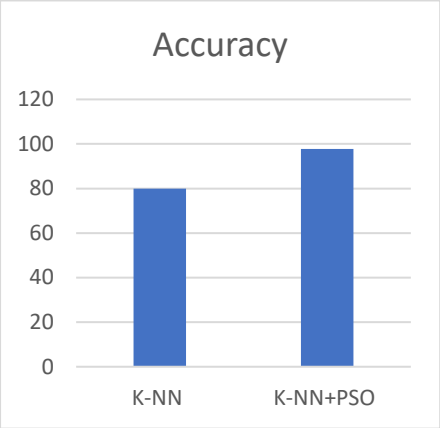

Figure 2. Comparison Diagram of Accuracy

In Figure 2, the comparison diagram of CKD data accuracy using K-NN and PSO shows that there is an increase in accuracy of 17.75% by integrating both methods. The results of both methods indicate that the dataset tested for the study using these methods performs better with a significant improvement over previous methods and is capable of correctly classifying the data.

CONCLUSION

This research has successfully demonstrated the significant potential of integrating the K-Nearest Neighbors (KNN) method with Particle Swarm Optimization (PSO) in improving the accuracy of Chronic Kidney Disease (CKD) prediction. As anticipated in the introduction, the results obtained confirm that the proposed approach is capable of addressing this issue by enhancing prediction accuracy up to 97.75%, a significant improvement compared to using KNN without optimization, which achieved an accuracy of 80.00%. The integration of KNN and PSO offers a more effective and efficient approach in identifying CKD at an early stage. This not only results in a more accurate predictive model but also enables dynamic parameter adjustments to adapt to the unique characteristics of the CKD dataset. Furthermore, this research also highlights the importance of feature selection through the PSO optimization process, where some attributes show significant contributions to the model's accuracy, while others do not provide meaningful contributions.

Therefore, the findings of this research signify an important advancement in CKD research and pave the way for further developments in this field. Prospects for developing these research findings include applying this method to larger and more diverse datasets to test its scalability and effectiveness under different conditions. Additionally, integration with technologies such as mobile applications and electronic health information systems can enhance the accessibility and usability of this predictive model, enabling broader early detection of CKD and timely medical interventions.

Further research could involve studying other optimization algorithms to see if further improvements in CKD prediction accuracy can be achieved. Moreover, further research can be conducted to understand the impact of implementing this model in daily clinical practice, including assessments of its acceptance by healthcare professionals and patients, as well as its impact on patient health outcomes. In conclusion, this research provides strong evidence that the integration of KNN and PSO is a highly promising approach in CKD prediction, offering new hope for early detection and effective management of this disease, which in turn can save more lives and reduce the economic burden of chronic kidney disease both in Indonesia and worldwide.

REFERENCES

Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease. *Journal of Information and Knowledge Management*, *19*(1). https://doi.org/10.1142/S0219649220400158

Arif-Ul-Islam, & Ripon, S. H. (2019). Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–6. https://doi.org/10.1109/ECACE.2019.8679388

Ariyati, I., Rosyida, S., Ramanda, K., Riyanto, V., Faizah, S., & Ridwansyah. (2020). Optimization of the Decision Tree Algorithm Used Particle Swarm Optimization in the Selection of Digital Payments. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012090

Iqbal, M., Herliawan, I., Ridwansyah, Gata, W., Hamid, A., Purnama, J. J., & Yudhistira. (2020). Implementation of Particle Swarm Optimization Based Machine Learning Algorithm for Student Performance Prediction. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *6*(2), 195–204. https://doi.org/10.33480/jitk.v6i2.1695.IMPLEMENTATION

Larose, D. T., & Larose, C. D. (2015). *Data Mining And Predictive Analytics*. John Wiley and Sons. *ISBN: 978-1-118-11619-7.*

Nurdin, H., Sartini, Sumarna, Maulana, Y. I., & Riyanto, V. (2023). Prediction of Student Graduation with the Neural Network Method Based on Particle Swarm Optimization. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(4), 2353–2362. https://doi.org/10.33395/sinkron.v8i4.12973

Pramanik, R., Khare, S., & Gourisaria, M. K. (2021). Inferring the Occurrence of Chronic Kidney Failure: A Data Mining Solution. *Proceedings of Second Doctoral Symposium on Computational Intelligence*. https://doi.org/https://doi.org/10.1007/978-981-16-3346-1_59

Purwaningsih, E. (2022). Improving the Performance of Support Vector Machine With Forward Selection for Prediction of Chronic Kidney Disease. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *8*(1), 18–24. https://doi.org/10.33480/jitk.v8i1.3327

Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*(December 2019), 100178. https://doi.org/10.1016/j.imu.2019.100178

Rezayi, S., Maghooli, K., & Saeedi, S. (2021). Applying Data Mining Approaches for Chronic Kidney Disease Diagnosis. *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*. https://doi.org/DOI: https://doi.org/10.18201/ijisae.2021473640

Ridwansyah, R., Riyanto, V., Hamid, A., Rahayu, S., & Purnama, J. J. (2022). Grouping Data in Predicting Infant Mortality Using K-Means and Decision Tree. *Paradigma*, *24*(2), 168–174. https://doi.org/10.31294/paradigma.v24i2.1399

Ridwansyah, R., Wijaya, G., & Purnama, J. J. (2020). Hybrid Optimization Method Based on Genetic Algorithm for Graduates Students. *Jurnal Pilar Nusa Mandiri*, *16*(1), 53–58. https://doi.org/10.33480/pilar.v16i1.1180

Saha, I., Gourisaria, M. K., & Harshvardhan, G. M. (2022). Classification System for Prediction of Chronic Kidney Disease Using Data Mining Techniques. *Lecture Notes in Networks and Systems*, *318*(May 2017), 429–443. https://doi.org/10.1007/978-981-16-5689-7_38

Sarker, I. H. (2021). *Machine Learning: Algorithms, Real-World Applications and Research Directions*. https://doi.org/https://doi.org/10.1007/s42979-021-00592-x.

Wijaya, G. (2024). *Improvement of Kernel SVM to Enhance Accuracy in Chronic Kidney Disease*. *9*(1), 136–144. https://doi.org/https://doi.org/10.33395/sinkron.v9i1.13112 e-ISSN

Witten, I. H. (2017). *DATA MINING (Fourth Edition)*. Elsevier. *ISBN: 9780128042915.*