NSS
Neutrosophic Sets and Systems, Vol. 80, 2025
University of New Mexico

Enhanced Neutrosophic Set and Machine Learning Approach for Kidney Disease Prediction

Humam M Al-Doori[1], Tareef S Alkellezli[2], Ahmed Abdelhafeez[3*,4], Mohamed Eassaa[3,4], Mohamed S. Sawah[5], Ahmed A El-Douh[3,6]

[1]Cybersecurity Sciences Department, College of Science, Ashur University, Baghdad, Iraq
[2]Cybersecurity Engineering Department, College of Engineering, Ashur University, Baghdad, Iraq
[3]Computer Science Department, Faculty of Information System and Computer Science, October 6 University, Giza, 12585, Egypt
[4]Applied Science Research Center. Applied Science Private University, Amman, Jordan
[5]Department of Computer Science, Faculty of Information Technology, Ajloun National University P.O.43, Ajloun-26810, Jordan
[6]Cybersecurity Technology Engineering Department, College of Engineering Technology, Ashur University, Baghdad, Iraq
*Correspondence: aahafeez.scis@o6u.edu.eg

Abstract:
Kidney disease (KD) is a gradually increasing global health concern. It is a chronic illness linked to higher rates of morbidity and mortality, a higher risk of cardiovascular disease and numerous other illnesses, and expensive medical expenses. The machine learning (ML) models are applied for KD prediction with higher accuracy and precision. The KD dataset has uncertainty and vague information, so, we used the neutrosophic set (NS) to deal with vague and uncertainty information in the KD dataset. The KD dataset is converted into the N-KD dataset with three membership functions: truth, indeterminacy, and falsity. Three ML models are used in this study such as logistic regression (LR), support vector machine (SVM), and k-nearest neighbor (KNN). These ML models are applied to the N-KD dataset. The results show the LR has higher accuracy and precision on the N-KD dataset than the original KD dataset.

Keywords: Neutrosophic Sets; Machine Learning Models; Kidney Disease; Uncertainty Models; Logistic Regression.

---

1. Introduction

Kidney disease (KD) is a gradually increasing global health concern. It is a chronic illness linked to higher rates of morbidity and mortality, a higher risk of cardiovascular disease and numerous other illnesses, and expensive medical expenses. Just 10% of people who require treatment to survive may be represented by the more than two million people who undergo dialysis or kidney transplants worldwide. Just five wealthy nations, which account for 12% of the world's population, are home to the majority of the two million renal failure patients who receive therapy. In contrast, barely 20% of the world's population is treated in the roughly 100 developing nations that make up about half of the world's population.[1], [2].

Due to the prohibitive cost of dialysis or kidney transplantation, over a million people in 112 lower-income countries pass away from untreated renal failure each year. The early identification, management, and control of the condition are therefore crucial.[3], [4]. Due to patient heterogeneity and the dynamic and hidden nature of KD in its initial stages, it is crucial to forecast its progression with a decent degree of precision. Stages of severity are frequently used to define KD. The stage, whether a patient is progressing, and the rate of progression all affect clinical judgments.[2], [5]. Determining the disease stage is also especially important because it provides several indicators that help determine the necessary interventions and therapies.

In the healthcare industry, machine learning (ML) algorithms have been employed for classification and prediction. The Support Vector Machine Algorithm (SVM) has been utilized by Yu et al. [6] to categorize and predict patients with diabetes and pre-diabetes. The findings indicate that SVM is helpful in classifying patients with common diseases. Like this, Magnin et al. [7] Used a Support Vector Machine (SVM) to classify Alzheimer's disease by analyzing whole-brain anatomical magnetic resonance imaging (MRI) for a group of patients. The findings indicate that SVM is a promising method for early Alzheimer's disease identification.[8], [9]. The KD dataset has uncaring information. This uncertainty has a negative effect on the ML models. So, in this study, we used the uncertainty framework to overcome the uncertainty data in the KD dataset.

1.1 Neutrosophic Sets

As an extension of intuitionistic and fuzzy logic, neutrosophic logic was presented. The expansion of fuzzy logic known as intuitionistic logic involves two degrees of membership: the degree of Falseness (non-membership) and the degree of Truth (membership). However, intuitionistic logic may not be enough to address the inherent inconsistency or indeterminacy levels that are frequently present in fuzzy systems because it can only accommodate incomplete information.[10], [11]. Neutrosophic sets, which are represented by degrees of membership for Truth, Indeterminacy, and Falseness, were developed to get around this limitation. The total of those three independent values falls between 0 and 3.

Based on a relatively recent area of philosophy called neutrosophic sets, these sets can simulate human knowledge, preferences, and evaluation schemes by addressing the inherent ambiguities, inconsistencies, and indeterminacies in each set of data.[12], [13].

Because natural systems are complex, decision-makers frequently face several uncertainties when making choices based on imprecise, unclear, and incomplete information. In recent years, the theory of fuzzy sets (FSs) and its numerous extensions have become popular approaches for dealing with incomplete data[14], [15]. However, the existing structures need to be better able to handle the ambiguous and inconsistent data that is frequently present in natural systems.

To alleviate this specific limitation, Smarandache put out the notion of Neutrosophic Sets (NSs). NS theory is a useful approach for handling data that has inconsistencies, ambiguity, and flaws. It can be thought of as an extension of the FS and intuitionistic fuzzy set. NSs offers a unique and useful paradigm for dealing with ambiguity and uncertainty in a variety of domains[16], [17]. The main benefit of NSs is their capacity to accurately portray real-world uncertainties by capturing a three-component representation: truth, indeterminacy, and falsehood.

1.2 Neutrosophic Set with ML Models

The NS can be used with ML models to improve accuracy and precision score. The NS can convert the original dataset into the neutrosophic dataset to overcome uncertainty and vague information in the training ML models.

Ahmed et al. [18] used the combination of NS with the ML models to reduce the subjectivity pertaining to expert driven decisions and produce a reliable ordering of the hospital building assets. Their model can aid in building consistent, unbiased and automated ranking for decision making issues.

Khan and Alghamdi [19] used the neutrosophic set to evaluate the IIoT devices trust score with the ML models. They used the neutrosophic clustering to classify the extracted features.

Kaya and Dengiz [20] used the neutrosophic set and ML models to detect soil quality index and then assess the performance of the ML models. They obtained higher performance based on neutrosophic set.

Saqlain et al. [21] used the neutrosophic set to solve the decision-making problems with the ML models. They used the ML models to rank the selection of sites for a new store.

Ashika et al. [22] used the neutrosophic set with the ML models for breast cancer prediction. They used the neutrosophic set for capturing uncertainty in original dataset. They converted their original dataset into a neutrosophic set.

The main contributions of this study are:

I. Converting the KD dataset into the N-KD dataset.
II. Training three ML models such as LR, KNN, and SVC.
III. Evaluating these ML models into accuracy, precision, recall, and f1 score.
IV. Comparing three ML models on the N-KD dataset and KD original dataset.

The rest of this study is organized as follows: Section 2 shows the materials and methods of this study with the NS to overcome the vague information. Section 3 shows the results and discussion of the three ML models on the N-KD dataset. Section 4 shows the comparison analysis between the three ML models on the N-KD dataset and the original dataset. Section 5 shows the conclusions of this study.

2. Materials and Methods

The materials and methods in this study are presented in this section by several steps such as:

2.1 Proposed methodology

The aim of this study is to present the ML models with NS for the prediction of kidney disease (KD). The dataset was gathered from the Kaggle website and then preprocessed to ensure its quality and consistency. The KD dataset is converted to the N-dataset and each data point is featured with three values instead of one value such as truth T, indeterminacy I, and falsity F. This approach can overcome the uncertainty and vague information on the KD dataset.

We can convert the KD dataset into the N-dataset such as:

$KD_N = f_{T,I,F}(KD)$

We can split the N-KD dataset into training and testing such as:

$KD_N = s(f_{KD_N}) = s(X_{train}, X_{test}, y_{train}, y_{test})$

We can normalize the N-KD dataset using the Standard Scaler method.

$KD_N = f_{Nor}\left(KD_{N_{(s)}}\right)$

The normalized N-KD dataset is trained using different ML models such as Logistic Regression, SVC, and KNN. These models are chosen based on their ability to deal with intricate features interacting and identify subtle patterns to obtain accurate results in KD prediction.

$$KD_{N_{(ML)}} = f_{ML}\left(KD_{N_{(Nor)}}\right) = KD_{ML}\left(LR_{KD_{N_{(Nor)}}}, SVC_{KD_{N_{(Nor)}}}, KNN_{KD_{N_{(Nor)}}}\right)$$

The ML models are validated using accuracy, precision, recall, and f1 score.

$$KD_{N_{(Metrics)}} = KD_{N(ML)}\left(Metrics_{Acc,Pre,Recall,F1\ score}\right)$$

Finally, this study performed a comparative analysis between the N-KD dataset and the original dataset to show the accuracy between them.

$$KD_{Comp} = Comp(KD_N, KD)$$

The flowchart of this study is shown in Figure 1.

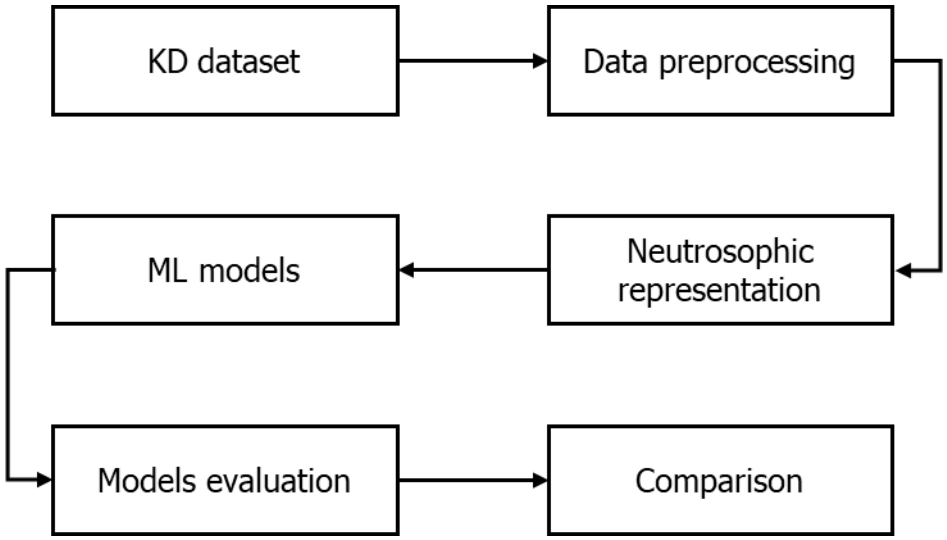

Figure 1. The flowchart of KD dataset prediction.

2.2 Dataset description

KD is used to predict kidney disease. It has four hundred rows and twenty-six columns. Table 1 shows the description of the KD dataset with some statistical methods to show the details of the dataset. Then we plot the distribution of the points in the dataset as shown in Figure 2. Figure 3 shows the box plot of each feature. Then we compute the count of each class in each feature as shown in Figure 4. Then we obtain the correlation matrix as shown in Figure 5.

We show the normal class has more rows than abnormal in the red blood cells. We show the normal class has more rows more than abnormal in the pus cells. We show the not present class has more rows more than present in the pus cell clumps cells. We show the present class has more rows than present in the bacteria cells. We show the no class has more rows than yes in the hypertension cells. We show the no class has more rows than yes in the diabetes mellitus cells. We show the no class has more rows than yes in the coronary artery disease cells. We show the good class has more rows than the poor in the appetite cells. We show the no class has more rows than yes in the pedal edema cells. We show the no class has more rows than yes in the anemia cells. We show the has disease class has more rows more than has not disease in the target class.

Table 1. Description of the dataset

	count	mean	std	Min	25%	50%	75%	Max
Id	400	199.5	115.6143	0	99.75	199.5	299.25	399
Age	391	51.48338	17.16971	2	42	55	64.5	90

blood pressure	388	76.46907	13.68364	50	70	80	80	180
specific gravity	353	1.017408	0.005717	1.005	1.01	1.02	1.02	1.025
Albumin	354	1.016949	1.352679	0	0	0	2	5
Sugar	351	0.450142	1.099191	0	0	0	0	5
blood_glucose_random	356	148.0365	79.28171	22	99	121	163	490
blood urea	381	57.42572	50.50301	1.5	27	42	66	391
serum creatinine	383	3.072454	5.741126	0.4	0.9	1.3	2.8	76
Sodium	313	137.5288	10.40875	4.5	135	138	142	163
Potassium	312	4.627244	3.193904	2.5	3.8	4.4	4.9	47
Hemoglobin	348	12.52644	2.912587	3.1	10.3	12.65	15	17.8
packed_cell_volume	329	38.8845	8.990105	9	32	40	45	54
white_blood_cell_count	294	8406.122	2944.474	2200	6500	8000	9800	26400
red_blood_cell_count	269	4.707435	1.025323	2.1	3.9	4.8	5.4	8

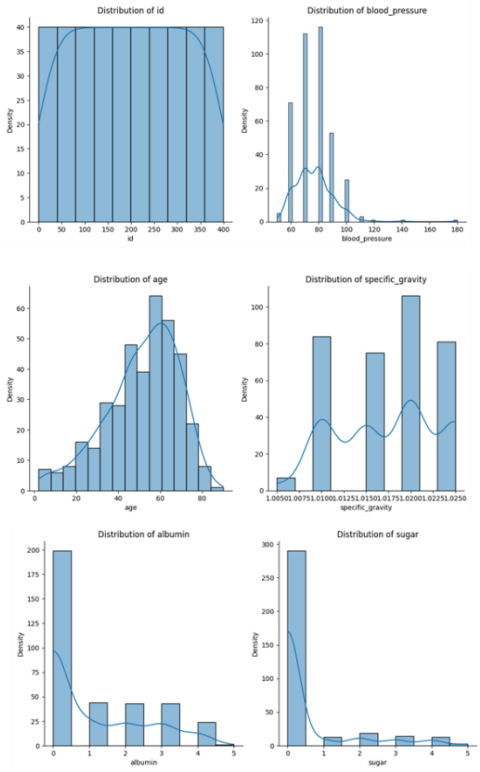

Figure 2. The distribution of the dataset.

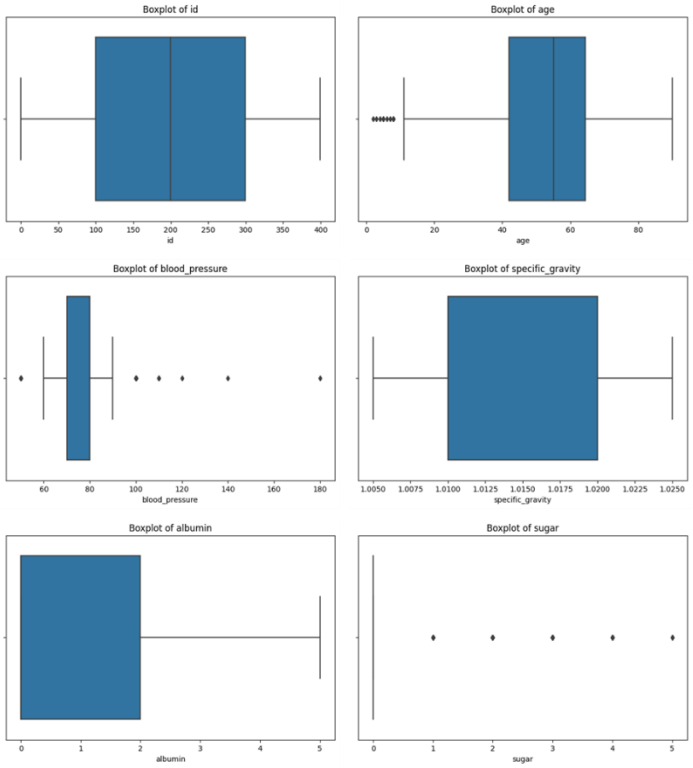

Figure 3. The box plot of the dataset.

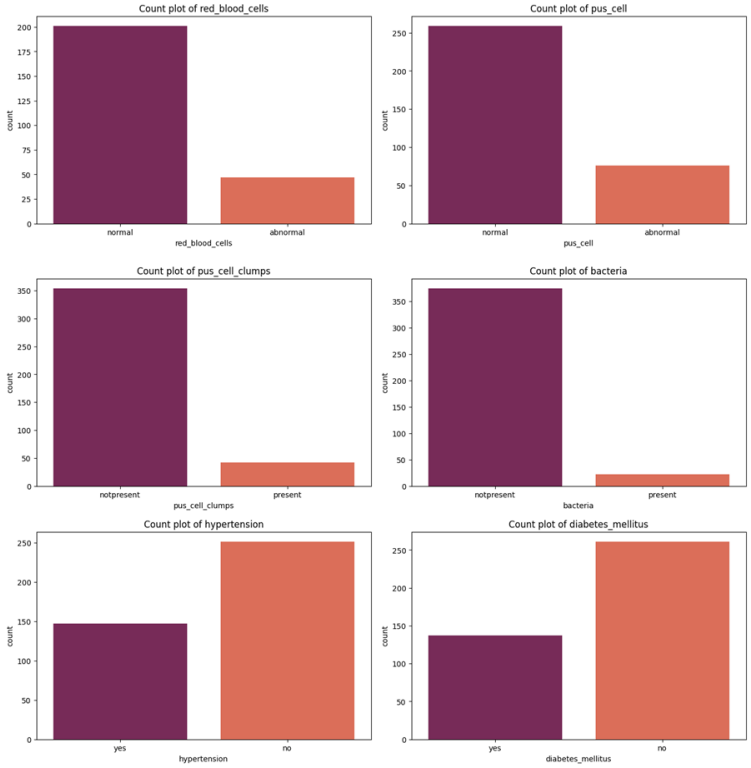

Figure 4. The Count of each point in each feature.

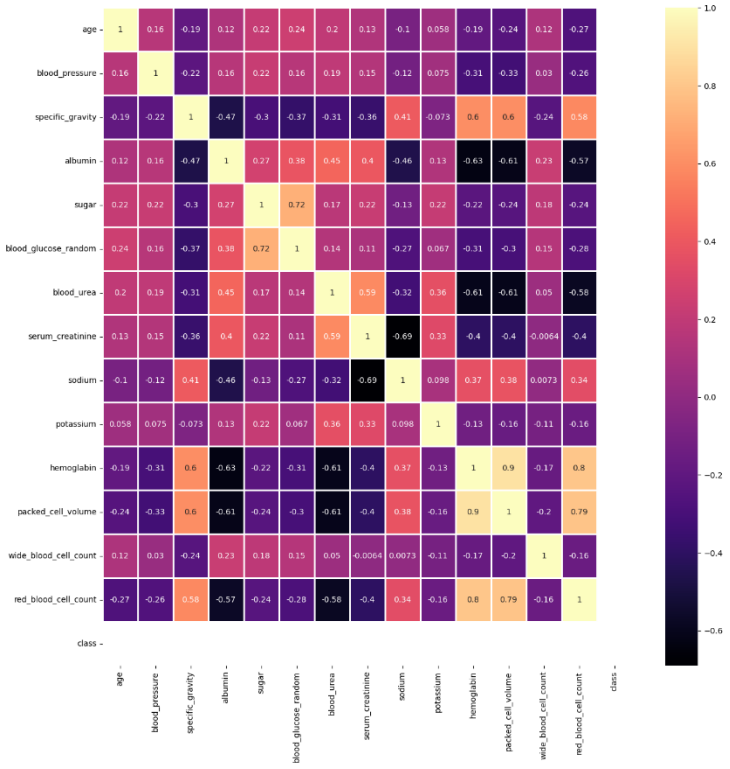

Figure 5. The Correlation matrix.

2.3 Data Preprocessing

The N-KD dataset was processed to obtain higher accuracy from the ML models. We divide the features with diseases and without diseases. Then this target class is encoded into 0 and 1. Then all features are normalized with the standard scaler method. These steps can improve the ML model evaluation on the N-KD dataset.

2.4 Neutrosophic Sets (NS)

The NS has three membership functions Truth, indeterminacy, and Falsity.[23], [24]. Where $X: T, I, F$ These values can be obtained with functions such as:

$$T_A(X), I_A(X), F_A(X)$$
$$T_A \; maps \; X \; to \; the \; interva \;]0^-, 1^+[$$
$$I_A \; maps \; X \; to \; the \; interva \;]0^-, 1^+[$$
$$F_A \; maps \; X \; to \; the \; interva \;]0^-, 1^+[$$

The sum of these functions is between 0 and 3. This flexibility can enable the NS to deal with uncertainty and vague data. The N-KD dataset contains the three elements of NS.

Definition 1

NS has three membership functions such as:

$$A_N: X \rightarrow [0^-, 1^+]$$
$$B_N: X \rightarrow [0^-, 1^+]$$
$$C_N: X \rightarrow [0^-, 1^+]$$
$$0^- \leq A_N(x) + B_N(x) + C_N(x) \leq 3^+$$

Definition 2

Let N and M are different NS. N can be considered a subset of M, if and only if conditions meet such as:

$$Inf\ A_N(x) \le Inf\ A_M(x)$$
$$Sup\ A_N(x) \le Sup\ A_M(x)$$

$$Inf\ B_N(x) \le Inf\ B_M(x)$$
$$Sup\ B_N(x) \le Sup\ B_M(x)$$

$$Inf\ C_N(x) \le Inf\ C_M(x)$$
$$Sup\ C_N(x) \le Sup\ C_M(x)$$

Definition 3

Considered two neutrosophic numbers such as $x = (A_1, B_1, C_1)\ and\ y = (A_2, B_2, C_2)$

$$x \oplus y = \begin{pmatrix} A_1(x) + A_2(x) - A_1(x)A_2(x) \\ B_1(x)B_2(x), \\ C_1(x)C_2(x) \end{pmatrix}$$

$$x \otimes y = \begin{pmatrix} A_1(x)A_2(x), \\ B_1(x) + B_2(x) - B_1(x)B_2(x), \\ C_1(x) + C_2(x) - C_1(x)C_2(x) \end{pmatrix}$$

$$\sigma x = \begin{pmatrix} 1 - \left(1 - A_1(x)\right)^{\sigma}, \\ B_1^{\sigma}(x), \\ C_1^{\sigma}(x) \end{pmatrix}$$

$$x^{\sigma} = \begin{pmatrix} A_1(x)^{\sigma}, \\ 1 - \left(1 - B_1(x)\right)^{\sigma}, \\ 1 - \left(1 - C_1(x)\right)^{\sigma} \end{pmatrix}$$

2.5 Neutrosophic Dataset Formation

To overcome the uncertainty in the KD dataset we convert it into the N-KD dataset.[22].

$X = \{x_1, x_2, \dots, x_n\}$

$\forall x(t, i, f) \in < T_A, I_A, F_A >$

We added the N-components into the original dataset to solve the vague and uncertainty information. We obtain the three functions such as:

$$T = 1 - \frac{\|x - U^+\|}{\max(\|x_{train} - U^+\|)}$$
$$T = 1 - \frac{\|x - U^{all}\|}{\max(\|x_{train} - U^{all}\|)}$$
$$T = 1 - \frac{\|x - U^-\|}{\max(\|x_{train} - U^-\|)}$$
$$U^+ = \sum_{n^+} x$$
$$U^{all} = \sum_{n^{all}} x$$
$$U^- = \sum_{n^-} x$$

2.6 ML Models

Logistic Regression (LR)

A supervised machine learning approach called logistic regression predicts the likelihood of an outcome, occurrence, or observation to complete binary classification tasks[25].

Different examples of classification problems

KNN

The k-nearest neighbor algorithm, or KNN, is a machine learning method that compares a single data point with a set of data it has learned and memorized to generate predictions based on proximity[26].

SVC

By carrying out optimal data transformations that establish boundaries between data points based on predefined classes, labels, or outputs, supervised learning models enable support vector machines (SVMs), a type of machine learning algorithm, to solve challenging classification, regression, and outlier detection problems [27].

3. Results and Discussion

This section shows the results of the proposed approach under the N-KD dataset.

3.1 Experimental Setup

Three ML models are used to predict the KD dataset. The dataset is converted to the N-KD dataset to improve the accuracy of the ML models and deal with uncertainty and vague information. Standard metrics are employed in this study to evaluate the ML models. We compared it with the original KD dataset to show the effectiveness of the NS in dealing with vague information.

3.2 Experimental Results

Figure 6 shows the results of the ML models under the N-KD dataset. We show the LR has higher accuracy, followed by the KNN, and SVC. We obtained accuracy, precision, recall, and f1 score.

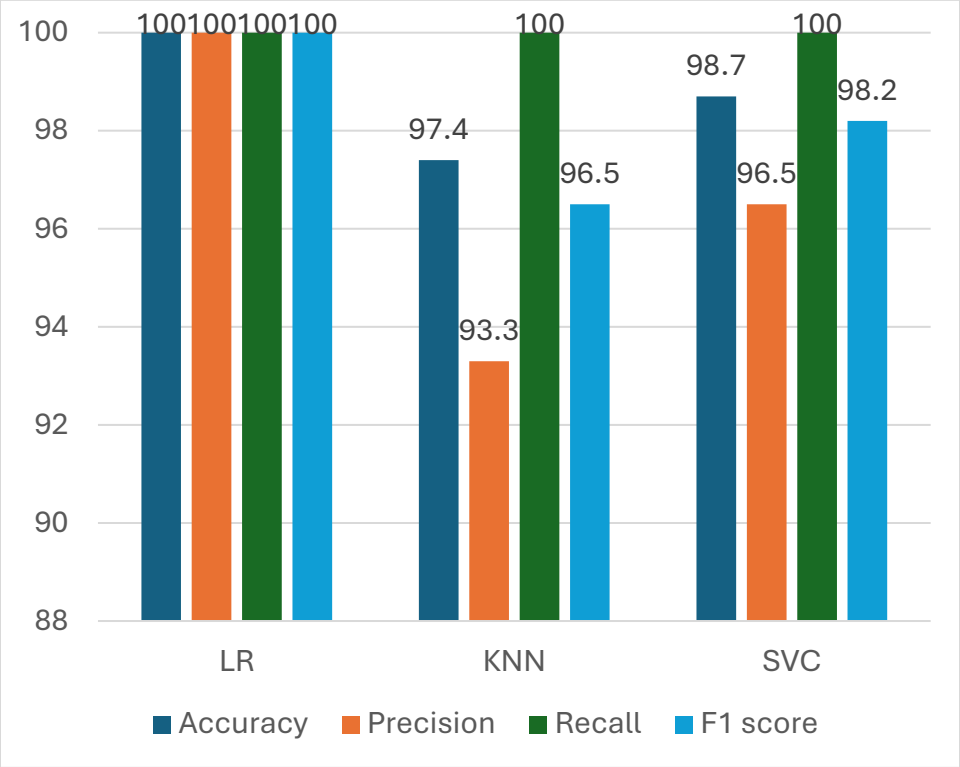

Figure 6. The ML models result under the N-KD dataset.

4. Comparison Analysis

We compare the ML models on the N-KD dataset and the original dataset. We show the accuracy, precision, recall, and f1 score as a classification report, as shown in Figure 7. Then, we obtained the confusion matrix

before and after the N-KD dataset, as shown in Figure 8. We show that the ML models obtained higher accuracy scores on the N-KD dataset than the original due to the NS and overcoming the uncertainty and vague information.

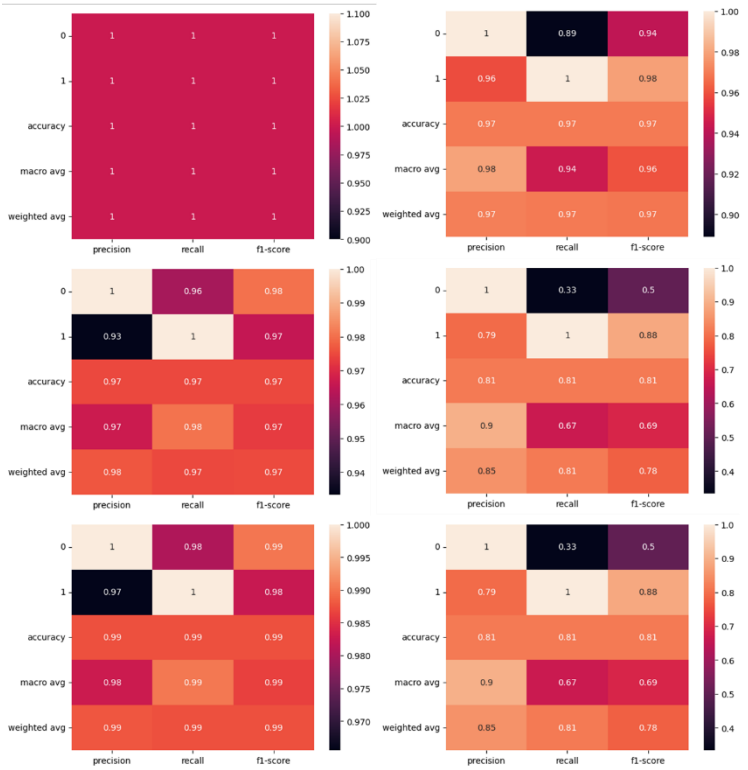

Figure 7. The classification report on the N-KD dataset and original dataset.

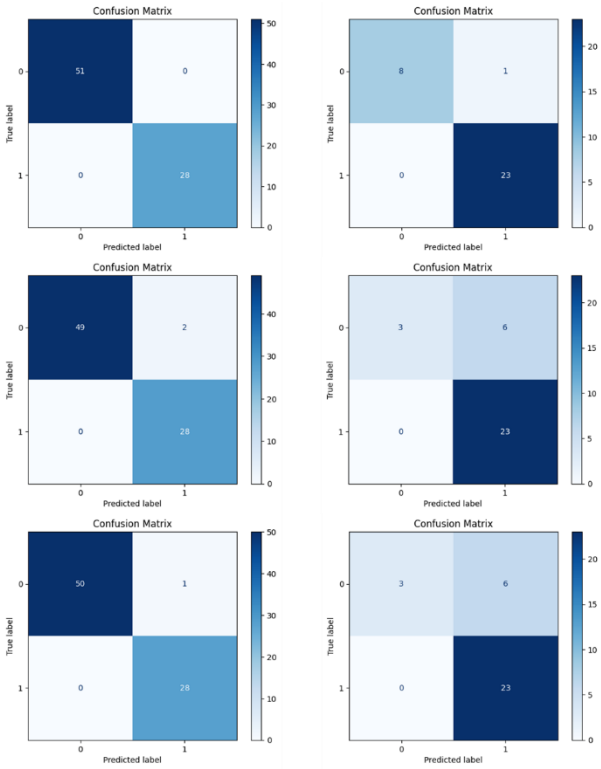

Figure 8. The confusion matrix on the N-KD dataset and original dataset.

5. Conclusions

This study integrated the ML models with the neutrosophic set to predict the KD dataset with uncertainty and vague data in the KD dataset. We trained three ML models such as LR, KNN, and SVC to predict the KD dataset. We converted the original KD dataset into the N-KD dataset with three values: truth, indeterminacy, and falsity. We applied the dataset preprocessing steps to improve the quality of the dataset. The results show the LR has higher accuracy than two other ML models on the N-KD dataset. We compare three ML models on the N-KD dataset and the original KD dataset. We show the ML models obtained higher accuracy than the original KD dataset. The three values of NS can deal with uncertainty and improve the accuracy of the ML models. We obtained the classification report between the ML models and the confusion matrix to show the true class and predicted class. In the future study, the proposed model can be applied to different prediction and classification problems to obtain higher accuracy and precision.

References

[1] W. G. Couser, G. Remuzzi, S. Mendis, and M. Tonelli, "The contribution of chronic kidney disease to the global burden of major noncommunicable diseases," *Kidney Int.*, vol. 80, no. 12, pp. 1258–1270, 2011.

[2] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics Med. unlocked*, vol. 15, p. 100178, 2019.

[3] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *J. Big Data*, vol. 9, no. 1, p. 109, 2022.

[4] P. Sinha and P. Sinha, "Comparative study of chronic kidney disease prediction using KNN and SVM," *Int. J. Eng. Res. Technol.*, vol. 4, no. 12, pp. 608–612, 2015.

[5] J. B. Echouffo-Tcheugui and A. P. Kengne, "Risk models to predict chronic kidney disease and its progression: a systematic review," *PLoS Med.*, vol. 9, no. 11, p. e1001344, 2012.
[6] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.*, vol. 10, pp. 1–7, 2010.
[7] B. Magnin *et al.*, "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," *Neuroradiology*, vol. 51, pp. 73–83, 2009.
[8] 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, no. 1, p. 6141470, 2021.
[9] N. Tangri *et al.*, "Risk prediction models for patients with chronic kidney disease: a systematic review," *Ann. Intern. Med.*, vol. 158, no. 8, pp. 596–603, 2013.
[10] F. Smarandache, *A unifying field in logics: neutrosophic logic. Neutrosophy, neutrosophic set, neutrosophic probability: neutrsophic logic. Neutrosophy, neutrosophic set, neutrosophic probability*. Infinite Study, 2005.
[11] S. Broumi, A. Bakali, and A. Bahnasse, "Neutrosophic sets: An overview," *Infin. Study*, 2018.
[12] A. A. Salama and S. A. Alblowi, "Neutrosophic set and neutrosophic topological spaces," 2012.
[13] N. El-Hefenawy, M. A. Metwally, Z. M. Ahmed, and I. M. El-Henawy, "A review on the applications of neutrosophic sets," *J. Comput. Theor. Nanosci.*, vol. 13, no. 1, pp. 936–944, 2016.
[14] A. M. Shitaya, M. E. S. Wahed, A. Ismail, M. Y. Shams, and A. A. Salama, "Predicting student behavior using a neutrosophic deep learning model," *Neutrosophic Sets Syst.*, vol. 76, pp. 288–310, 2025.
[15] M. Sharma, I. Kandasamy, and W. B. Vasantha, "Comparison of neutrosophic approach to various deep learning models for sentiment analysis," *Knowledge-Based Syst.*, vol. 223, p. 107058, 2021.
[16] A. Z. M. Elsherif *et al.*, "Unveiling Big Data Insights: A Neutrosophic Classification Approach for Enhanced Prediction with Machine Learning," *Neutrosophic Sets Syst.*, vol. 72, pp. 154–172, 2024.
[17] A. Nafei, S. P. Azizi, S. A. Edalatpanah, and C.-Y. Huang, "Smart TOPSIS: a neural Network-Driven TOPSIS with neutrosophic triplets for green Supplier selection in sustainable manufacturing," *Expert Syst. Appl.*, vol. 255, p. 124744, 2024.
[18] R. Ahmed, F. Nasiri, and T. Zayed, "A novel Neutrosophic-based machine learning approach for maintenance prioritization in healthcare facilities," *J. Build. Eng.*, vol. 42, p. 102480, 2021.
[19] M. A. Khan and N. S. Alghamdi, "A neutrosophic WPM-based machine learning model for device trust in industrial internet of things," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 4, pp. 3003–3017, 2023.
[20] N. S. Kaya and O. Dengiz, "Assessment of the neutrosophic Fuzzy-AHP and predictive power of some machine learning approaches for maize silage soil quality," *Comput. Electron. Agric.*, vol. 226, p. 109446, 2024.
[21] M. Saqlain, H. Garg, P. Kumam, and W. Kumam, "Uncertainty and decision-making with multi-polar interval-valued neutrosophic hypersoft set: A distance, similarity measure and machine learning approach," *Alexandria Eng. J.*, vol. 84, pp. 323–332, 2023.
[22] H. Grace, N. Martin, and F. Smarandache, "Enhanced Neutrosophic Set and Machine Learning Approach for Breast Cancer Prediction," *Neutrosophic Sets Syst.*, vol. 73, no. 1, p. 20, 2024.
[23] A. Elhassouny, S. Idbrahim, and F. Smarandache, "Machine learning in neutrosophic environment: a survey," *Infin. study*, 2019.
[24] M. Alshikho, M. Jdid, and S. Broumi, "Artificial Intelligence and Neutrosophic Machine learning in the Diagnosis and Detection of COVID 19," *J. Prospect. Appl. Math. Data Anal.*, vol. 1, no. 2, 2023.
[25] T. G. Nick and K. M. Campbell, "Logistic regression," *Top. Biostat.*, pp. 273–301, 2007.
[26] S. Zhang, "Challenges in KNN classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp.

4663–4675, 2021.
[27] M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1194–1203, 2007.

Received: Sep 5, 2024. Accepted: Dec 31, 2024