

A comparative study of machine learning methods for drug type classification

Andi Tejawati¹, Didit Suprihanto², Aji Ery Burhandenny³, Saipul⁴, Novianti Puspitasari¹, Anindita Septiarini¹

¹Department of Informatics, Faculty of Engineering, Mulawarman University, Samarinda, Indonesia

²Department of Electronic Engineering, Faculty of Engineering, Mulawarman University, Samarinda, Indonesia

³Department of Electrical Engineering Education, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

⁴Department of Communication Sciences, Faculty of Social and Political Sciences, Mulawarman University, Samarinda, Indonesia

Article Info

Article history:

Received Oct 25, 2024

Revised May 10, 2025

Accepted May 27, 2025

Keywords:

Classification

Cross-validation

Drug types

K-nearest neighbor

Machine learning

ABSTRACT

Drugs, commonly called narcotics, are dangerous substances that, if consumed excessively, can result in addiction and even death. Drug abuse in Indonesia has reached a concerning stage. In 2017, the National Narcotics Agency detected 46,537 drug-related incidents, including methamphetamine, marijuana, and ecstasy. There are 4 types of substances that can affect drug users, such as hallucinogens, depressants, opioids, and stimulants. A machine learning approach can detect these substances using user symptom data as input. This study uses six different methods in classifying, including decision tree, C.45, K-nearest neighbor (KNN), random forest, and support vector machine (SVM). The dataset comprises 144 data and 21 attributes based on the user's symptoms. The evaluation method in this study uses cross-validation with K-fold values of 5 and 10 and uses three parameters: precision, recall, and accuracy. KNN yields the most optimal results by using K=1 and K-fold 10 in the Euclidean and Minkowski types. The model achieves precision, recall, and accuracy of 91.9%, 91.7%, and 91.67%, respectively.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Novianti Puspitasari

Department of Informatics, Faculty of Engineering, Mulawarman University

Sambaliung Street No. 9, Samarinda, Indonesia

Email: novia.ftik.unmul@gmail.com

1. INTRODUCTION

Social and environmental criteria have exerted an impact on others to participate in drug consumption. Adult individuals who have a substance addiction exert a significant effect on the conduct of others towards developing addiction. Inadequate abilities, high levels of stress, anxiety, and aberrant behavior are other elements that contribute to drug use [1]. Pharmacological substances, regardless of their origin as natural, synthetic, or semi-synthetic, possess the capacity to cause changes in consciousness, hallucinations, and sedation. The phenomena of drug addiction are sometimes classified into four identifiable stages: sporadic use, recreational use, chronic use, and severe addiction. The progression of these stages is significantly shaped by emotions, consciousness, and cognition [2]. Substance abuse is a major public health issue causing physiological symptoms, behavioral changes, cognitive impairments, and mental health issues. It impacts future generations and leads to addiction or relapse. Global drug misuse is a widespread epidemic, causing severe physical and psychological damage. Due to the high prevalence of drug users, scientific studies on drugs have gained significant attention. The number of drug users from year to year has increased throughout the years 2022-2023, recording 4.8 million people as drug users. However, the classification of drug users and information about the type and remarks of drug users in East Kalimantan Province is still unknown. Therefore,

a classification is needed. In recent years, machine learning has been a potent tool in the media world. Several studies in the medical field that have implemented machine learning include chronic kidney [3], [4], cardiovascular [5], [6], Alzheimer's [7], brain tumor [8], breast cancer [9], prediction of covid 19 [10], and drugs field [11]. Types of data commonly used in implementation using machine learning in the health field include signals [12], images [13], or medical record data consisting of names, ages, laboratory test results, and disease symptoms [14], [15]. While the machine learning method that is commonly used in the medical world is generally such as logistic regression and random forest [16], [17], artificial neural network (ANN) [18], Naïve Bayes [19], support vector machine (SVM) [20], and K-nearest neighbor (KNN) [21], [22].

The National Narcotics Agency in Indonesia has been actively combating drug abuse despite the country's growing vulnerability to illicit trafficking. In 2017, the agency detected 46,537 drug-related incidents, including methamphetamine, marijuana, and ecstasy. However, the country's lack of oversight and the province of East Kalimantan's inability to categorize drug users further highlight the need for a system to identify and determine the level of drug users in the country. By utilizing machine learning techniques and medical data, research in the medical field has been widely applied to several studies. An ANN with the backpropagation algorithm categorizes dengue types into DF, DHF, and DSS. A dataset of 21 dengue symptoms from 110 patients was used. Cross-validation with K-fold 2, 3, 5, and 10 was used for evaluation. The best performance was achieved with K-fold 3 cross-validation, achieving precision, recall, and accuracy values of 0.969, 0.967, and 0.967, respectively [18]. An intelligent fuzzy system was proposed to diagnose and predict the risk of developing type 2 diabetes mellitus. The system consists of two models: the R-T2DM model, which estimates risk, and the DT2DM model, which estimates symptoms and diagnoses. The R-T2DM model achieved a success rate of 90.3%, while the D-T2DM model achieved 88.3% and 95.5% success rates. The model is designed for use in economically marginalized Mexico areas to improve patient quality of life [23].

An algorithm to classify was proposed using nine data types, including eight kinds of narcotic data acquired from the above portable IMS detector and under general conditions. Various types of narcotics include amphetamine, morphine hydrochloride, fentanyl, alfentanil hydrochloride, MDMA hydrochloride, ketamine hydrochloride, diazepam (Dia), and codeine phosphate hydrate. The proposed system algorithm for detecting inferred drugs from narcotic IMS data achieved average accuracies of 0.9 for KNN, 0.94 for TSF, and 0.99 for ROCKET, confirming its good performance [24]. The survival outcomes involve real-world data, particularly for oncology patients. PUBMED and EMBASE were searched for peer-reviewed English-language studies on ML models for predicting time-to-event outcomes using RWD, extracting data sources, patient population, survival outcome, ML algorithms, and AUC. The study included 28 publications out of 257 citations, with random survival forests and neural networks being the most popular machine learning algorithms. These models were primarily used for predicting overall survival in oncology (N=12, 43%), disease prognosis or clinical events (N=27, 96%), and treatment outcomes (N=1, 4%). Variability across AUC was observed [25].

This work aims to classify drug types into four categories: hallucinogens, depressants, opioids, and stimulants. The input data was derived from the symptoms caused by the drug usage. The classification was performed using several machine-learning methods, including Naive Bayes, SVM, KNN, C.45, random forest, and decision tree. The dataset was divided into train and test sets using cross-validation. The structure of the paper is as follows: section 2 outlines the dataset and methods, section 3 presents the results and discussion of each classification method, and section 4 provides the conclusion.

2. METHOD

The following part overviews the dataset details and the classification approach employed. Furthermore, it offers details on the procedure for assessing the effectiveness of each classification technique. The information used in this study was obtained from National Narcotics Agency, East Kalimantan, Indonesia, and comprised 144 instances of drug addiction. The dataset is partitioned into four categories, namely hallucinogens, depletes, opioids and stimulants, representing 44, 39, 45, and 16 data points, respectively. The data was collected in the form of drug user codes, age and symptoms experienced. Each user may experience different symptoms, leading to a variation in the diagnosis of drug users by experts. The dataset consisted of 21 types of symptoms experienced by users from medical record data for the period 2018–2020. Table 1 displays examples of data collected from drug users.

The age parameter in Table 1 has no significant effect because the initial analysis indicates a weak correlation with the type of drug used. Variations across age groups in the dataset do not show a consistent or significant pattern that would improve the classification model's accuracy. Therefore, including this parameter would only increase complexity without meaningfully contributing to the model's performance. This study focuses on more relevant features that directly impact drug type classification, so age is excluded to maintain the model's efficiency and effectiveness.

Table 1. Examples of medical record data of drug users

User code	Age	Symptoms	Diagnose
A1	43	Increased heart rate, red eyes, closed eyelids close, often anxious and panicked, insomnia, widening pupils, decreased appetite, and slow reflexes.	Hallucinogens
A2	24	Slow reflexes, concentrations are disturbed, often anxious and panicked, easily drowsy, dizziness, and headache	Deples
A3	19	Red eyes, easy to laugh, concentration disturbed, decreased appetite, and hallucinations	Hallucinogens
A4	28	Disturbed concentration, often anxious and panicked, easy to forget, easy to sleep, and nausea	Deples
A5	30	Insomnia, hallucinations, nausea, and dry mouth	Hallucinogens
A6	29	Decreased appetite, muscle pain, dizziness or headache, confusion, and red eyes	Opioids
:	:	:	:
A139	23	Dizziness or headache, hallucinations, slow reflexes, and muscle aches	Stimulant
A140	32	Pupils wide, decreased appetite, hallucinations, and red eyes	Stimulant
A141	23	Red eyes, eyelids close, easy to laugh, concentration disturbed, dizziness or headaches, and decreased appetite	Hallucinogens
A142	27	Easy to forget, slow reflexes, nausea, vomiting, and easy to sleep	Deples
A143	15	Decreased appetite, nausea, vomiting, pupil shrinking, and red eyes	Opioids
A144	29	Often anxious and panicked, dizziness or headache, nausea, vomiting, and easy to forget	Deples

The present study comprises two distinct phases, namely training and testing. The two primary operational processes executed by both systems are pre-processing and categorizing. Furthermore, it is crucial to establish a systematic evaluation procedure to quantify the efficiency of particular classifiers. The assessment methodology utilizes the diagnosis acquired from the expert system (an actual class) and the classification technique (a predicted class) as input. Figure 1 illustrates the overview of the dengue classification approach.

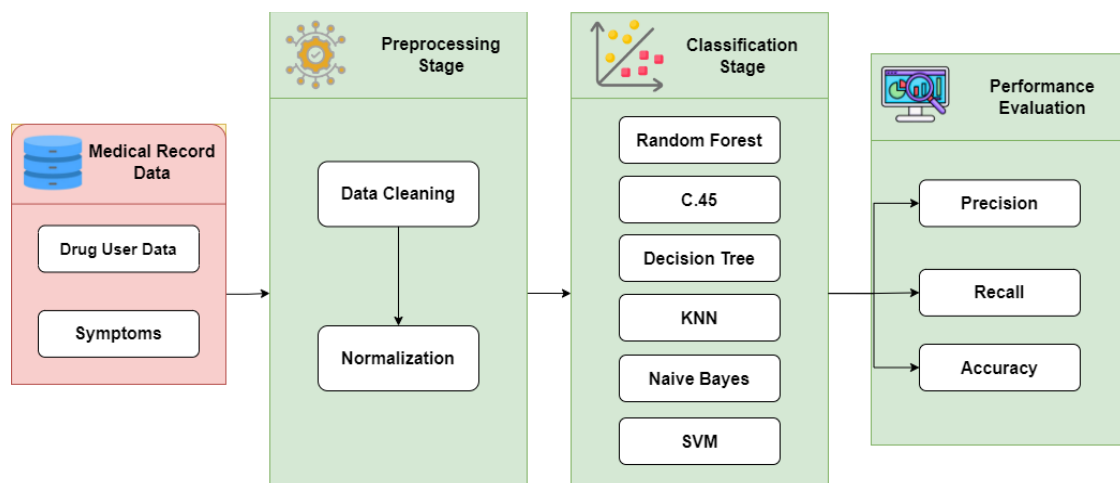


Figure 1. The overview of the processes on the drug usage level classification method

2.1. Pre-processing

Discretization was applied as a pre-processing step. The data shown in Table 1 needed to be converted into numerical format to serve as input for the classification process. The drug user data contained 21 distinct symptom levels (G), such as increased heart rate (G01), red eye (G02), drooping eyelids (G03), excessive appetite (G04), easily laughing (G05), impaired concentration (G06), frequent anxiety and panic (G07), insomnia (G08), dilated pupils (G09), headache (G10), decreased appetite (G11), hallucinations (G12), fatigue (G13), slow reflexes (G14), forgetfulness (G15), nausea (G16), vomiting (G17), constricted pupils (G18), dry mouth (G19), muscle pain (G20), and drowsiness (G21). Thus, these 21 values were inputs for the subsequent classification process. The symptoms reported by drug users in Table 1 are represented as categorical variables, with 1 indicating the presence of a symptom and 0 indicating its absence. Meanwhile, age and drug user code data did not require pre-processing. The gathered data is prepared for use in the classification process through pre-processing. The statistical data pre-processing for this study is shown in Table 2.

Table 2. The result of pre-processing medical record data

Code	G1	G2	G3	G4	G5	G6	G7	:	G16	G17	G18	G19	G20	G21	Diagnose
A1	1	1	1	0	0	0	1	:	0	0	0	0	0	0	Hallucinogens
A2	0	0	0	0	0	1	1	:	0	0	0	0	0	1	Deples
A3	1	0	0	0	0	0	0	:	1	0	0	0	0	0	Hallucinogens
A4	0	0	0	0	0	0	0	:	1	1	1	0	0	0	Deples
A5	1	0	0	0	0	1	0	:	1	1	0	0	0	1	Hallucinogens
A6	0	0	0	0	0	0	1	:	0	0	0	0	0	0	Opioids
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
A139	0	0	0	0	0	0	0	:	1	0	0	0	0	0	Stimulant
A140	0	0	0	0	0	0	0	:	1	1	0	0	0	0	Stimulant
A141	0	0	0	1	0	0	0	:	0	0	0	1	0	0	Hallucinogens
A142	0	0	0	0	0	1	0	:	0	0	0	0	0	1	Deples
A143	0	0	0	0	0	0	1	:	0	0	0	1	1	1	Opioids
A144	0	0	0	0	0	0	0	:	1	1	0	0	0	1	Deples

2.2. Classification method

Furthermore, after completing data cleansing for processing, we proceeded to the classification stage. This stage involved using six classification algorithms: Naive Bayes, decision tree, C.45, random forest, KNN, and SVM. These methods were chosen based on their successful implementation in a previous study [26] which served as the standard for this case study. The primary aim of this research was to explore the KNN approach for further classification of drug users. KNN was selected due to its strong, simple, and efficient performance in handling complex data, its versatility across different case studies, and its ability to manage imbalanced data, having been successfully applied in several studies [3], [5], [7]. A detailed explanation of the KNN classifier is provided in the following.

The KNN algorithm classifies data by utilizing the training data from the k nearest neighbors, where k represents the number of nearest neighbors considered. KNN performs classification in multiple dimensions using projected learning data. The training data points exist in a multi-dimensional space. The KNN method requires a positive integer value, k , to define the number of neighbors used for the classification task. The newly classified data is then projected into this multi-dimensional space. Classification is carried out by identifying the closest point. Several distance formulas are provided in (1)-(3) [27]:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (2)$$

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad (3)$$

where N is the number of attributes, x is the data test, and y is the data train, while 1 is the Manhattan distance formula, 2 is the Manhattan distance, and 3 is the Minkowski distance. The p -value in the formula can be manipulated to give another distance, like $p=1$ Manhattan distance or $p=2$ Euclidean distance. The KNN model in this study was built based on a dataset from a pre-processing medical data record.

2.3. Performance evaluation

The performance of the classification approach was evaluated using three metrics: precision (p), recall (r), and accuracy (a), which were based on the confusion matrix multiclass structure. The evaluation parameters have a numerical value ranging from 0 to 100. The approach performs sufficiently when those parameters approach a value of 100. The following parameters are specified (4)-(6) [3]:

$$p = \frac{\sum_{k=1}^n N_{ki}}{N_{ii}} \times 100, \quad (4)$$

$$r = \frac{\sum_{k=1}^n N_{ki}}{N_{ii}} \times 100, \quad (5)$$

$$a = \frac{\sum_{i=1}^n N_{ii}}{\sum_{i=1}^n \sum_{j=1}^n N_{ij}} \times 100, \quad (6)$$

The dataset is divided into k subsets according to the specific type of diagnostic data. The dataset $(k-1)/k$ is allocated for training, whereas the dataset $1/k$ is reserved for testing. The procedure is subsequently repeated K -fold. The final rate estimate is established by choosing the validation result of the mean k -time as the concluding point. This work assesses performance through cross-validation employing K -fold values of 5 and 10.

3. RESULTS AND DISCUSSION

Pre-processing is performed using discrete methods to convert all drug user data into numerical formats. Based on user data, 21 symptoms (G1 to G21) have been identified. Thus, 21 characteristics were employed as input data for the subsequent procedure, precisely that of classification. A value of 1 was assigned to symptom data if the user experienced it. Otherwise, its value is null if the user did not encounter it. In the classification of drug users, there are four distinct categories: hallucinogens, depletes, opioids, and stimulants. The classification approach involves using C.45, decision tree, KNN, random forest, Naive Bayes, and SVM since these classifiers have been successfully applied to solve various cases. This approach aimed to achieve the most efficient performance method, which was evaluated using three specific metrics: precision, recall, and accuracy. This number was derived from a multiclass confusion matrix calculated using a cross-validation method with three distinct K-fold values: 5 and 10. An analysis of the performance of six classification methods is presented in Table 3.

Table 3. Drug type classification results using various classifiers and K-fold

Classifier	K-fold 5			K-fold 10		
	<i>p</i> (%)	<i>r</i> (%)	<i>a</i> (%)	<i>p</i> (%)	<i>r</i> (%)	<i>a</i> (%)
Naive Bayes	77.2	77.1	77.08	78.70	78.50	78.47
SVM	86.80	86.80	86.80	84.50	84	84.02
KNN	87.80	87.50	87.50	91.90	91.70	91.67
C.45	72	71.50	71.52	73.40	73.60	73.61
Random forest	90.1	89.6	89.58	90.20	90.30	90.27
Decision tree	70	66	65.97	70.20	71.50	71.52

The classification results for the drug usage levels using various classifiers—Naive Bayes, SVM, KNN, C4.5, random forest, and decision tree are presented in Table 3. The performance metrics evaluated include precision (*p*), recall (*r*), and accuracy (*a*) across two different K-fold validation setups: K-fold 5 and K-fold 10. Among all the classifiers, random forest demonstrated the best overall performance, namely achieving a precision of 90.1%, recall of 89.6%, and accuracy of 89.58% for the 5-fold validation. The 10-fold configuration displayed a similarly high precision of 90.20%, recall of 90.30%, and accuracy of 90.27%. These results demonstrate the resilience of the random forest algorithm in effectively managing intricate information such as medical records. However, KNN also performed exceptionally well, achieving stable and competitive results. In the 5-fold cross-validation, KNN attained 87.80% precision, 87.50% recall, and 87.50% accuracy.

During the 10-fold validation, the performance of the KNN algorithm improved even further, with precision reaching 91.90%, recall at 91.70%, and accuracy at 91.67%. Both validation methods consistently highlighted the robustness of KNN in accurately classifying drug use levels. Naive Bayes, SVM, and C4.5 among the classifiers tested produced similar results. The Naive Bayes model achieved a precision of 77.2% in the 5-fold test, while the SVM model showed an accuracy of 86.80%. KNN demonstrated remarkable consistency and accuracy, particularly in the 10-fold validation. In contrast, the decision tree and C4.5 models exhibited lower accuracy compared to KNN, with the Decision Tree model specifically attaining accuracy between 65.97% and 71.52%.

Table 4 displays the results of the KNN classifier implemented with Euclidean, Minkowski, and Manhattan distance metrics and *K* values of 1, 3, and 5 derived from 5-fold and 10-fold cross-validation. Once again, precision (*p*), recall (*r*), and accuracy (*a*) are emphasized as the primary performance measures. With 87.8% precision, 87.5% recall, and 87.5% accuracy in the 5-fold cross-validation and 91.9% precision, 91.7% recall, and 91.67% accuracy in the 10-fold setting, *K*=1 achieved the highest performance for the Euclidean distance metric in both validation scenarios. However, performance significantly decreased as the value of *K* increased to 3 and 5. At *K*=5, for instance, the precision dropped to 74.1% and accuracy to 73.61% in the 5-fold setup, and similar declines were seen in the 10-fold validation.

Table 4. The classification result using different distance algorithm and K value

Distance	K	K-fold 5			K-fold 10		
		<i>p</i> (%)	<i>r</i> (%)	<i>a</i> (%)	<i>p</i> (%)	<i>r</i> (%)	<i>a</i> (%)
Euclidean	1	87.8	87.5	87.5	91.9	91.7	91.67
	3	75.9	75	75	75.2	74.3	74.3
	5	74.1	73.6	73.61	76.9	76.4	76.3
Minkowski	1	87.8	87.5	87.5	91.9	91.7	91.67
	3	75.9	75	75	75.2	74.3	74.3
	5	74.1	73.6	73.61	76.9	76.4	76.3
Manhattan	1	87.8	87.5	87.5	91.9	91.7	91.67
	3	75.9	75	75	75.2	74.3	74.3
	5	74.1	73.6	73.61	76.9	76.4	76.3

The results for the Minkowski distance were similar to those of the Euclidean distance, with $K=1$ consistently yielding the best performance across both cross-validation techniques. Minkowski's precision, recall, and accuracy at $K=1$ matched those of Euclidean, indicating that both distance metrics perform similarly for this dataset when K is set to 1. As with Euclidean, performance declined as K values increased, particularly at $K=5$, where precision and accuracy dropped to 76.9% and 76.3% in the 10-fold validation. Although following the same trend, Manhattan distance produced slightly lower results than Euclidean and Minkowski. At $K=1$, it still achieved a precision of 87.8% in the 5-fold cross-validation. However, $K=5$ showed the weakest performance among the three-distance metrics, with precision at 74.1% and accuracy at 73.61% in the 5-fold setup, with further decline in the 10-fold evaluation. Figure 2(a) illustrates the confusion matrix for the classification results using KNN with a K -fold value of 5; meanwhile, Figure 2(b) applies a K -fold value of 10.

TARGET \ OUTPUT	Hallucinogens	Deples	Opiods	Stimulan	SUM
Hallucinogens	37 26.69%	1 0.69%	4 2.78%	2 1.39%	44 84.09% 15.91%
Deples	0 0.00%	37 25.59%	1 0.69%	1 0.69%	39 94.87% 5.13%
Opiods	1 0.69%	3 2.08%	40 27.78%	1 0.69%	45 88.89% 11.11%
Stimulan	1 0.69%	0 0.00%	3 2.08%	12 8.33%	16 75.00% 25.00%
SUM	39 94.87% 5.13%	41 90.24% 9.76%	48 83.33% 16.67%	16 75.00% 25.00%	126 / 144 87.50% 12.50%

(a)

TARGET \ OUTPUT	Hallucinogens	Deples	Opiods	Stimulan	SUM
Hallucinogens	40 27.78%	1 0.69%	0 0.00%	3 2.08%	44 90.91% 9.09%
Deples	0 0.00%	37 25.59%	1 0.69%	1 0.69%	39 94.87% 5.13%
Opiods	1 0.69%	0 0.00%	43 29.86%	1 0.69%	45 95.56% 4.44%
Stimulan	1 0.69%	0 0.00%	3 2.08%	12 8.33%	16 75.00% 25.00%
SUM	42 95.24% 4.76%	38 97.37% 2.63%	47 91.49% 8.51%	17 70.59% 29.41%	132 / 144 91.67% 8.33%

(b)

Figure 2. The confusion matrix of the classification result using KNN with; (a) K -fold=5 and (b) K -fold=10

The findings indicate that $K=1$ is the most effective value across all distance metrics, though increasing K -values diminishes classification accuracy. Euclidean and Minkowski distance metrics outperform Manhattan, especially at lower K values, making them better for classifying drug usage. This study reaffirms KNN as a robust classifier due to its stability with Euclidean and Minkowski distances, particularly in 5-fold and 10-fold cross-validations. While comparison research shows that random forest slightly outperforms KNN, the latter is more consistent and stable, justifying its selection as the main focus of this study.

4. CONCLUSION

Drugs are dangerous substances because they can affect nerve performance. Abuse of drugs can cause hallucinations and even lead to death. The existence of a system capable of classifying drug users can aid health professionals in quickly diagnosing users based on their symptoms. In this study, the types of substances that affect users include hallucinogens, depressants, opioids, and stimulants. This study found that KNN was the best classifier for putting drug use levels into groups based on medical records. It did better than or as well as random forest, SVM, C4.5, decision tree, and Naive Bayes. KNN had superior stability and performance, especially with Euclidean and Minkowski distance metrics with $K=1$, achieving precision, recall, and accuracy of 91.9%, 91.7%, and 91.67%, now 10-fold cross-validation. Random forest had slightly high metrics in some circumstances but maintained consistent findings across validation setups, making it more dependable for classification. The study also stressed the significance of choosing the correct k value, as greater values resulted in a significant performance reduction. This research emphasizes the KNN classifier is the most balanced and effective due to its simplicity, robustness, and ability to adapt to medical diagnostic data. This paper recommends that further research by researchers can increase the number of symptoms, parameters, and other algorithms that do not yet exist and can also be applied to a system for diagnosing the level of drug addiction.

ACKNOWLEDGMENTS

Thanks to National Narcotics Agency, East Kalimantan, Indonesia for their assistance in this research.

FUNDING INFORMATION

The research was funded by the Faculty of Engineering at Mulawarman University in Samarinda, Indonesia (Grant No. 8962/UN17.9/PT.00.03/2024) in 2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andi Tejawati	✓								✓			✓		✓
Didit Suprihanto	✓	✓					✓			✓				
Aji Ery Burhandenny Saipul			✓	✓		✓		✓	✓		✓			
Novianti Puspitasari		✓			✓		✓			✓				
Anindita Septiarini				✓	✓					✓				

- C : Conceptualization
M : Methodology
So : Software
Va : Validation
Fo : Formal analysis
- I : Investigation
R : Resources
D : Data Curation
O : Writing - Original Draft
E : Writing - Review & Editing
- Vi : Visualization
Su : Supervision
P : Project administration
Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, NP, upon reasonable request.

REFERENCES

[1] A. Anggrawan, N. G. A. Dasriani, Mayadi, C. Satria, C. K. Nuraini, and Lusiana, "Machine Learning for Diagnosing Drug Users and Types of Drugs Used," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 111–118, 2021, doi: 10.14569/IJACSA.2021.01211113.

[2] N. Basuni and A. M. Siregar, "Comparison of the Accuracy of Drug User Classification Models Using Machine Learning Methods," *J. RESTI (Rekayasa Sist. dan Teknol. Informatika)*, vol. 7, no. 6, pp. 1348–1353, 2023, doi: 10.29207/resti.v7i6.5401.

[3] R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa, and A. Alkhayyat, "Chronic Kidney Disease Detection Using Machine Learning Technique," in *IICETA 2022 - 5th Int. Conf. Eng. Technol. its Appl.*, 2022, pp. 153–158, doi: 10.1109/IICETA54559.2022.9888564.

[4] M. M. Hossain *et al.*, "Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease," *Mach. Learn. with Appl.*, vol. 9, p. 100330, 2022, doi: 10.1016/j.mlwa.2022.100330.

[5] S. Sravani and P. R. Karthikeyan, "Detection of cardiovascular disease using KNN in comparison with naive bayes to measure precision, recall and f-score," *AIP Conference Proceedings*, vol. 2821, no. 1, 2023, doi: 10.1063/5.0177014.

[6] S. J. Pasha and E. S. Mohamed, "Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction," *Informatics Med. Unlocked*, vol. 32, p. 101064, 2022, doi: 10.1016/j.imu.2022.101064.

[7] K. C. Reddy and S. Nithyaselvakumari, "Alzheimer's disease detection using cosine KNN classifier machine learning algorithm in comparison with medium KNN classifier with improved accuracy," *AIP Conference Proceedings*, vol. 2822, no. 1, 2023, doi: 10.1063/5.0173204.

[8] I. B. Santoso and S. N. Utama, "Multi-Model of Convolutional Neural Networks for Brain Tumor Classification in Magnetic Resonance Imaging Images," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, 2024, doi: 10.22266/IJIES2024.1031.56.

[9] M. Carrilero-Mardones, M. Parras-Jurado, A. Nogales, J. Pérez-Martín, and F. J. Díez, "Deep Learning for Describing Breast Ultrasound Images with BI-RADS Terms," *J. Imaging Informatics Med.*, no. 0123456789, 2024, doi: 10.1007/s10278-024-01155-1.

[10] S. Dasgupta, S. Das, and D. Chakraborty, "Prediction equations for detecting COVID-19 infection using basic laboratory parameters," *J. Fam. Med. Prim. Care*, vol. 13, no. 7, p. 2683, 2024, doi: 10.4103/jfmpe.jfmpe.

[11] G. M. Dimitri and P. Lió, "DrugClust: A machine learning approach for drugs side effects prediction," *Comput. Biol. Chem.*, vol. 68, pp. 204–210, 2017, doi: 10.1016/j.compbiolchem.2017.03.008.

[12] A. S. R. and H. C. Nagaraj, "Classification of EEG signal using EACA based approach at SSVEP-BCI," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, p. 717, 2021, doi: 10.11591/ijai.v10.i3.pp717-726.




[13] M. A. J. M. Kani, M. S. Parvathy, S. M. Banu, and M. S. A. Kareem, "Classification of skin lesion images using modified Inception V3 model with transfer learning and augmentation techniques," *J. Intell. Fuzzy Syst.*, vol. 44, no. 3, pp. 4627–4641, 2023.

[14] G. Feretzakis *et al.*, "Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy," *Healthc. Inform. Res.*, vol. 27, no. 3, pp. 214–221, 2021, doi:




- 10.4258/hir.2021.27.3.214.
- [15] A. Novianto and M. D. Anasanti, "Autism spectrum disorder (ASD) identification using feature-based machine learning classification model," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 17, no. 3, pp. 259–270, 2023.
 - [16] R. Gangula, L. Thirupathi, R. Parupati, K. Sreeveda, and S. Gattoju, "Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns," *Mater. Today Proc.*, no. 40, 2021, doi: 10.1016/j.matpr.2021.07.270.
 - [17] N. Innab *et al.*, "AI-Driven Predictive Modeling for Lung Cancer Detection and Management Using Synthetic Data Augmentation and Random Forest Classifier," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, pp. 1–20, 2025, doi: 10.1007/s44196-025-00879-4.
 - [18] H. Hamdani, Z. Arifin, and A. Septiari, "Expert System of Dengue Disease Using Artificial Neural Network Classifier," *JUITA J. Inform.*, vol. 10, no. 1, p. 59, 2022, doi: 10.30595/juita.v10i1.12476.
 - [19] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
 - [20] E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New Cardiovascular Disease Prediction Approach Using Support Vector Machine and Quantum-Behaved Particle Swarm optimization," *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 23901–23928, 2024.
 - [21] Z. E. Fitri, L. N. Y. Syahputri, and A. M. N. Imron, "Classification of White Blood Cell Abnormalities for Early Detection of Myeloproliferative Neoplasms Syndrome Based on K-Nearest Neighbor," *Sci. J. Informatics*, vol. 7, no. 1, pp. 136–142, 2020, doi: 10.15294/sji.v7i1.24372.
 - [22] A. Septiari, D. M. Khairina, A. H. Kridalaksana, and H. Hamdani, "Automatic glaucoma detection method applying a statistical approach to fundus images," *Healthc. Inform. Res.*, vol. 24, no. 1, pp. 53–60, 2018, doi: 10.4258/hir.2018.24.1.53.
 - [23] J. R. Grande-Ramírez, R. Meza-Palacios, A. A. Aguilar-Lasserre, R. Flores-Asis, and C. F. Vázquez-Rodríguez, "Intelligent fuzzy system to assess the risk of type 2 diabetes and diagnosis in marginalized regions," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, p. 1935, 2024, doi: 10.11591/ijai.v13.i2.pp1935-1944.
 - [24] S. Park, G. Kemelbekova, S. Cho, K. Kwon, and T. Im, "Study on the Ion Mobility Spectrometry Data Classification and Application of Port Container Narcotics Using Machine Learning Algorithm," *Appl. Sci.*, vol. 13, no. 23, 2023, doi: 10.3390/app132312769.
 - [25] Y. Huang, J. Li, M. Li, and R. R. Aparasu, "Application of machine learning in predicting survival outcomes involving real-world data: a scoping review," *BMC Med. Res. Methodol.*, vol. 23, no. 1, pp. 1–11, 2023, doi: 10.1186/s12874-023-02078-1.
 - [26] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," *Procedia Comput. Sci.*, vol. 218, pp. 249–261, 2023, doi: 10.1016/j.procs.2023.01.007.
 - [27] A. Salam, Sri S. Prasetyowati, and Y. Sibaroni, "Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest," *J. RESTI (Rekayasa Sist. dan Teknol. Informatika)*, vol. 4, no. 3, pp. 531–536, 2020, doi: 10.29207/resti.v4i3.1926.

BIOGRAPHIES OF AUTHORS






Andi Tejawati    is a lecturer at the at the Department of Informatics, Mulawarman University. She is attached to the Computing and Informatics Institutions Indonesia (APTIKOM) societies. Her research interests include social informatics and artificial intelligence. She can be contacted at email: anditejawati.ifunmul@gmail.com.






Dedit Suprihanto    holds a Bachelor of Engineering in Informatics Engineering, Master of Information System, Ph.D. in Department of Computer Science and Electronics Gadjah Mada University, besides several professional certificates and skills. He is currently lecturing with the Department of Electronics Engineering at Mulawarman University, Samarinda, Indonesia. He is a member of the Indonesian Computer, Electronics, and Instrumentation Support Society (IndoCEISS). His research areas of include computer networks security, e-Government related issues and security assessment. He can be contacted at email: dedit.suprihanto@ft.unmul.ac.id.






Aji Ery Burhandenny    is an assistant professor in Department of Electrical Engineering Education, Universitas Negeri Yogyakarta, Indonesia, specializing in Software Engineering. His research interest lie in empirical software engineering particularly understanding human factors in software development activities. He also made several contributions to internet of things, machine learning, and green technologies related topics. He can be contacted at email: ajieryburhandenny@uny.ac.id.






Saipul    is a lecturer at the Department of Communication Sciences, Faculty of Social and Political Sciences, Mulawarman University. His research interest includes social problems and communication technology. He can be contacted at email: andialmer25@yahoo.com.



Novianti Puspitasari    received the B.Sc. degree in informatics engineering from the Universitas Islam Indonesia, and the M.Eng. degree in information technology from the Gadjah Mada University, Indonesia. She is currently a lecturer at the Department of Informatics, Mulawarman University. She is a member of the Institute of Electrical and Electronics Engineers (IEEE), Indonesian Computer, Electronics, Instrumentation Support Society (IndoCEISS), Association of Computing and Informatics Institutions Indonesia (APTIKOM) societies and The Institution of Engineers Indonesia (PII). She has authored or coauthored more than 70 publications with 5 H-index. Her research interest is in data science and analytics, artificial intelligence, and machine learning areas. She can be contacted at email: novia.ftik.unmul@gmail.com.



Anindita Septiarini    is a professor at the Department of Informatics at Mulawarman University, Indonesia. She holds a Doctoral degree in Computer Science from Gadjah Mada University, Indonesia, specializing in image analysis. She is also a researcher and got a grant from the Ministry of Education, Culture, Research, and Technology of Indonesia from 2016 until the present. Her research interests lie in artificial intelligence, especially pattern recognition, image processing, and computer vision. She has received national awards such as scientific article incentives from the Ministry of Education, Culture, Research, and Technology of Indonesia in 2017 and 2019. She held several administrative posts with the Department of Informatics, Mulawarman University, Indonesia, from 2018 to 2020, including the Head of Department and the Head of Laboratory. She can be contacted at email: anindita@unmul.ac.id.