Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i1.1013 Published By DRPM-UBD

A Comparative Study of Drug Prediction Models using KNN, SVM, and Random Forest

Susi Eva Maria Purba

Information System Departement, Institut Teknologi Del, Toba, Indonesia
Email: susipurba2@gmail.com

Abstract

Accurate drug classification is essential in medical decision-making to ensure patients receive appropriate prescriptions based on their physiological and biochemical characteristics. This study compares the performance of K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest models in predicting drug prescriptions using patient attributes such as age, sex, blood pressure, cholesterol level, and sodium-to-potassium ratio. The dataset, obtained from Kaggle, was preprocessed and split into training and testing sets to evaluate model performance using accuracy as the primary metric. The results indicate that Random Forest outperformed KNN and SVM, achieving a perfect test accuracy of 100%, demonstrating superior generalization and robustness. SVM also performed well, with a test accuracy of 97.50%, while KNN achieved the lowest accuracy of 70%, indicating its limitations in handling complex feature interactions. These findings highlight the effectiveness of ensemble learning methods in medical classification tasks, suggesting that Random Forest is the most suitable model for drug prediction. Furthermore, the potential applications of these findings in clinical settings could enhance treatment outcomes and patient care. Future research should explore feature engineering techniques, larger datasets, and additional machine learning approaches to enhance predictive accuracy and applicability in real-world healthcare settings.

Keywords: Drug classification, Machine Learning, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest

1. INTRODUCTION

Essential medical diagnosis and treatment programs rely heavily on patient data. Using this data correctly allows for precise forecasts, specific treatment plans, and better healthcare results [1]. In recent years, artificial intelligence (AI) and machine learning (ML) have transformed healthcare through offering advanced tools for analyzing large amounts of data. These technologies are essential for forecasting the effectiveness of drugs, suggesting appropriate prescriptions, and reducing side effects. Healthcare professionals can improve decision-making and lessen the workload related to conventional techniques by using AI-driven technologies [1], [2].

While patient data is critical in medical diagnosis and treatment, the growing complexity and volume of healthcare datasets provide substantial hurdles to reliable analysis and decision-making. Predicting the best medicine for a patient requires examining a variety of data types, including categorical factors such as gender and cholesterol levels, as well as numerical variables such as age and sodium-to-potassium ratio. Traditional methods often have a hard time handling different types of data well, which can result in poor predictions.

Machine learning (ML) models provide a solution by automating the study of large information and detecting patterns that might otherwise go undiscovered. However, finding the most effective ML model remains a crucial difficulty due to algorithm performance variability [1], [3], [4]. The promise of models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest have been highlighted. Previous studies have explored various machine learning algorithms for drug prediction, emphasizing their potential applications and limitations [5], [6], [7], [8]. These studies have shown the strengths of each model in various circumstances. However, their effectiveness in handling datasets with mixed categorical and numerical features for drug prediction remains an open question. The challenge lies in how different algorithms process categorical and numerical data, affecting their ability to generalize well in real-world medical applications. Addressing this gap is essential for improving automated drug prescription systems, potentially leading to more personalized and effective treatments.

This study compares the performance of three commonly used medication prediction models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. These models were chosen for their distinct strengths in dealing with diverse datasets: KNN for its ease of use and effectiveness in classifying data based on similarity [9], SVM for its ability to find optimal decision boundaries in complex and non-linear datasets [10], and Random Forest for its robustness and high accuracy in capturing complex feature relationships [11], [12]. This study evaluates these models on a dataset containing both numerical (e.g., Age, Na to K) and categorical (e.g., Sex, Blood Pressure, Cholesterol) characteristics. Age, Na-K ratio, sex, blood pressure, and cholesterol are critical health indicators that directly impact treatment decisions. For instance, age affects drug metabolism and response, the Na to K ratio is linked to electrolyte balance and blood pressure regulation, and physiological differences by sex can influence drug efficacy. Likewise, blood pressure and cholesterol levels serve as essential cardiovascular health markers that guide prescription choices.

By systematically comparing these models, this study seeks to identify the most effective approach for drug prediction in datasets containing mixed data types, addressing a key challenge in machine learning applications for healthcare [13],

[14]. The results will help identify the best model for predicting drug groups, tackling issues with varied medical data. The rest of this paper is organized as follows: Section 2 explains the dataset, preprocessing, machine learning models, and evaluation metrics. Section 3 presents and analyzes the performance of the models. Section 4 summarizes findings and suggests future research directions. Section 5 lists the cited works supporting this study.

2. METHODS

The methods section involves understanding the business problem to define objectives and guide data collection. It includes exploring and describing the data to uncover patterns, verifying its quality, and preparing it for analysis through data cleaning, transformation, and encoding. Finally, various models are applied and evaluated to address the business objectives, with statistical insights used to interpret and optimize the results. The process followed in this study is illustrated in Figure 1.

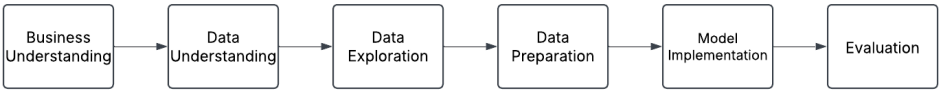

Figure 1. Methodology

2.1. Business Understanding

The main aim of this study is to review and compare the prediction abilities of three machine learning models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—in classifying drug types based on patient attributes. Effective drug classification is crucial in medical decision-making, as it ensures that patients receive appropriate prescriptions based on their physiological and biochemical characteristics. By analyzing various patient-related factors, this research aims to identify the most accurate and generalizable model for drug prediction, providing insights into the applicability of machine learning in healthcare.

2.2. Data Understanding

The dataset utilized in this study was obtained from Kaggle, a widely recognized open-source data repository. It contains structured patient information, including demographic details and biochemical markers, which are essential for drug classification. The dataset consists of five independent variables—Age, Sex, Blood Pressure (BP), Cholesterol, and Sodium-to-Potassium (Na to K) ratio—used to predict the dependent variable, Drug.

1) Age: A continuous variable representing the patient's age.

2) Sex: A categorical variable indicating the patient's gender (Male/Female).
3) Blood Pressure (BP): A categorical variable with three levels—Low, Normal, and High.
4) Cholesterol: A categorical variable with two levels—Normal and High.
5) Sodium-to-Potassium Ratio (Na to K): A continuous variable representing the ratio of sodium concentration to potassium concentration in the patient's system.
6) Drug: A categorical variable representing the prescribed drug class. This serves as the target variable.

Understanding the relationships among these variables is fundamental to developing an effective predictive model, as variations in these attributes influence drug prescriptions. Describing the data involves summarizing and characterizing the collected dataset to gain an initial understanding of its key features. A summary of the distribution and core patterns of the data may be obtained by calculating descriptive statistics like range, standard deviation, median, and mean. Histograms, bar charts, and box plots help visualize data and find patterns and trends. This step helps to identify any potential issues with the data and serves as the foundation for more in-depth analysis in the next stages of the project.

2.3. Data Exploration

Exploratory analysis was conducted to assess data distribution, identify potential outliers, and detect missing values. Descriptive statistics were utilized to summarize numerical features, and frequency distributions were analyzed for categorical variables. Box plots and histograms were employed to visualize data dispersion, particularly for Age and Na to K ratio, while bar charts were used to examine categorical feature distributions. Correlation analysis was performed to assess relationships between numerical variables, while cross-tabulation was used to explore interactions between categorical predictors and drug classifications. This preliminary analysis provided insights into feature importance and potential patterns in the data.

2.4. Data Preparation

Data preparation is a critical phase that involves cleaning and transforming the data into a format suitable for modeling. In the data preparation phase, relevant features were selected based on their potential influence on drug classification. The dataset was cleaned and preprocessed, ensuring consistency in categorical variable encoding and numerical feature normalization where necessary. In this step, data issues such as missing values, duplicates, or incorrect entries are addressed [15], [16]. Missing data can be handled using techniques such as imputation or removing rows/columns with too many missing values. Data transformation might include normalization or scaling of numerical features, encoding categorical variables, and

creating new features based on domain knowledge [16]. The target variable, Drug, was encoded into numerical representations for compatibility with machine learning models. The final dataset structure included:

1) Target variable: Drug classification.
2) Predictor variables: Age (continuous), Sex (categorical), BP (categorical), Cholesterol (categorical), and Na to K ratio (continuous).

Categorical variables such as Sex, BP, and Cholesterol were encoded using one-hot encoding, while numerical variables were scaled to maintain feature uniformity across models. The dataset was then split into training (80%) and testing (20%) subsets to evaluate model generalization.

2.5. Model

The modeling phase in machine learning involves exploring different algorithms to determine the most effective one for solving a given problem. In the context of classification and regression tasks, three common models often compared for their effectiveness are K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. Each of these algorithms has unique strengths and weaknesses, making them suitable for different types of data and problem structures.

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning method. It classifies new data points using the majority class or average value of the feature space's k-nearest neighbors. The user-defined "k" number specifies how many neighbors are evaluated while predicting [9], [16]. Support Vector Machine (SVM) is a strong supervised learning technique that is commonly used for classification problems. One advantage of SVM is that it provides a regularization hyperparameter C that, if properly adjusted, prevents overfitting [17]. SVM works by identifying a hyperplane that optimally separates various classes of data with the greatest feasible margin, hence increasing the model's generalizability.

Random Forest is an ensemble learning technique that builds many decision trees and then combines their outputs to get a final prediction. Each tree is created on a random selection of data and characteristics, which reduces the danger of overfitting as compared to individual decision trees. The Random Forest model combines predictions from all of the trees in the forest, usually using a majority vote for classification tasks or an average for regression. One of the primary benefits of Random Forest is its resistance to overfitting, especially when dealing with noisy data and big feature sets. Furthermore, Random Forest can handle both classification and regression problems and produces feature significance scores, which can be valuable in understanding the contribution of different variables to the model's predictions [11], [12], [16].

For instance, KNN may be appropriate for smaller datasets with non-linear decision boundaries, SVM may excel in high-dimensional spaces with clear class separations, and Random Forest may be the best choice when handling large datasets with complex relationships and a need for robustness. Three machine learning models KNN, SVM, and Random Forest were implemented and trained on the prepared dataset.

1) K-Nearest Neighbors (KNN): The KNN model was configured with specific hyperparameter settings to optimize its performance for the given dataset. These choices represent a balance between model complexity, sensitivity to local patterns, and robustness to noise.
    a) n_neighbors = 10:
    The choice of 10 neighbors is a trade-off between bias and variance. A smaller value (e.g., 1 or 2) can make the model overly sensitive to noise, increasing variance and leading to overfitting. A larger value averages predictions over more data points, which may smooth decision boundaries and cause underfitting (higher bias) [18]. A value of 10 was chosen as a balance, capturing local patterns while avoiding excessive sensitivity to noise or outliers.
    b) p = 1 (Manhattan Distance):
    The choice of Manhattan distance (L1 norm) over Euclidean distance (L2 norm) is intentional, especially when the features are not necessarily continuous or when data has a sparse or grid-like structure. Manhattan distance works well in situations where each dimension contributes independently to the distance, and this metric tends to be more effective in high-dimensional [19].
    c) spaces.weights = 'distance':
    Using distance-based weighting assigns greater importance to closer neighbors when making predictions [20]. This is useful in cases where the data exhibits non-linear relationships or varying densities. It allows the model to focus on more relevant, nearby points and reduces the influence of distant points, which may be less relevant, especially in high-dimensional spaces where the curse of dimensionality can make most points appear similarly distant.
2) Support Vector Machine (SVM): The SVM model was trained with specific hyperparameter settings to optimize its performance and generalization capabilities. The choices made reflect a balance between model complexity, computational efficiency, and the characteristics of the data.
    a) kernel = ‘linear’:
    The linear kernel was chosen because it assumes the data is linearly separable, which is computationally efficient and simpler to interpret. If the data were highly complex or non-linearly separable, a more complex kernel (like radial basis function) would be more appropriate. A linear kernel is often a good first choice for classification tasks,

especially when the feature space is not expected to require complex transformations [21].

b) C = 1:
The parameter C controls the trade-off between achieving a low error on the training data and minimizing the complexity of the decision boundary [22]. A value of 1 is a typical starting point, balancing model accuracy and generalization. If C were too large, the model would prioritize low training error, potentially leading to overfitting; if it were too small, the model might underfit by being too simple.

c) degree = 1:
This parameter controls the degree of the polynomial kernel, but since the linear kernel is used, it has no effect here. It's included as part of the default configuration for the sake of completeness. If a polynomial kernel were used, a degree of 1 would represent a linear decision boundary [21].

d) gamma = 0.01:
Gamma controls the influence of individual training points on the decision boundary [23]. A lower value (like 0.01) ensures that the model is less sensitive to individual data points and can generalize better. In this case, a small gamma prevents the model from overfitting by making it less sensitive to noise or outliers.

3) Random Forest: The Random Forest model was trained with specific hyperparameter configurations to optimize its predictive performance and computational efficiency. These choices reflect a balance between model accuracy, complexity, and training time.

a) criterion = 'gini':
The Gini impurity is a widely used metric for classification in decision trees, as it measures the "impurity" of a set of data. The lower the Gini index, the more homogenous the class labels are within a node. This criterion was chosen because it is efficient and works well in practice, particularly for binary or multi-class classification problems. It's a good balance between simplicity and effectiveness [24], [25].

b) n_estimators = 200:
Random Forest is an ensemble method, and the number of estimators (trees) plays a crucial role in its performance [26]. While adding more trees typically improves performance by reducing variance and avoiding overfitting, it also increases computational cost [26], [27]. A value of 200 is a reasonable choice that provides good model accuracy while maintaining computational efficiency. More trees beyond this point generally show diminishing returns in accuracy but increase training time.

Each model was trained on the training set and evaluated on the test set, with performance measured using accuracy scores. These scores were compared to assessing the models' predictive efficacy and determine the most reliable classification method for drug prescription.

2.6. Metric Evaluation

Once the model has been constructed and trained, it is necessary to assess its performance using pertinent metrics. The performance of the K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest models were evaluated primarily on the basis of accuracy. Out of all the predictions, accuracy is the percentage of properly categorized cases. It is defined as shown in Equation 1.

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \quad (1)$$

Accuracy is a simple but effective measure of model performance that shows how effectively each model generalizes to previously encountered data. Training accuracy shows how well the model fits the training data, while test accuracy shows how well it can predict on unseen data; these metrics were used to evaluate the models [16].

3. RESULTS AND DISCUSSION

3.1. Experimental Performance

The results of this study highlight the varying effectiveness of different machine learning models—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—in predicting drug classification based on patient attributes. Table 1 presents the comparative performance of these models, measured in terms of training and test accuracy.

Tabel 1. Train vs Test Score based on Model

Model	Train Score (%)	Test Score (%)
KNN	76.25	70
SVM	98.75	97.50
Random Forest	99.38	100

The findings indicate that Random Forest achieved the highest predictive accuracy, with a training score of 99.38% and a perfect generalization score of 100% on the test set. In contrast, SVM displayed strong performance, attaining a training score of 98.75% and a test score of 97.50%, while KNN showed moderate predictive ability, with a training accuracy of 76.25% and a test accuracy of 70%.

The KNN model, implemented with n_neighbors=10, p=1, and weight=distance, exhibited moderate learning capability but struggled with generalization, as indicated by the gap between its training and test scores. This limitation stems from KNN's reliance on distance-based classification, making it highly sensitive to variations in feature scales. Additionally, its effectiveness diminishes when dealing with datasets containing both categorical and continuous features, leading to suboptimal decision boundaries in complex classification problems.

The SVM model, configured with C=1, degree=1, gamma=0.01, and a linear kernel, demonstrated significantly better performance than KNN, with a high degree of generalization. The linear kernel effectively captured the influence of categorical attributes like blood pressure (BP) and cholesterol while efficiently integrating the sodium-to-potassium (Na/K) ratio as a continuous feature. However, despite its high accuracy, SVM remains computationally demanding and requires careful parameter tuning, particularly in selecting the optimal kernel function for the dataset. These challenges must be considered when deploying SVM in practical applications.

Among the three models, Random Forest emerged as the most effective, surpassing both KNN and SVM in predictive accuracy. Using criterion=gini and n_estimators=200, it achieved near-perfect performance, demonstrating robust generalization ability. The ensemble learning nature of Random Forest allows it to efficiently handle complex interactions between patient attributes, such as the correlation between blood pressure, cholesterol levels, and the Na/K ratio, which play a crucial role in determining drug suitability. Moreover, Random Forest provides feature importance rankings, offering valuable insights into the key factors influencing drug classification decisions. This interpretability makes it particularly advantageous in medical applications, where understanding the reasoning behind a model's predictions is critical.

The superior performance of Random Forest can be attributed to its ability to mitigate overfitting by aggregating multiple decision trees, ensuring high predictive accuracy even when applied to unseen data. Additionally, its ability to capture non-linear relationships between categorical and numerical features enhances its reliability in drug classification tasks. However, computational efficiency remains a consideration, as the ensemble-based structure of Random Forest requires more processing power compared to simpler models like KNN.

Given these results, several optimization strategies could further improve the models' performance. Fine-tuning Random Forest's hyperparameters, such as n_estimators, max_depth, and min_samples_split, could enhance its accuracy while balancing computational efficiency. Feature engineering may also improve

predictions by incorporating additional patient-specific variables, such as lifestyle habits and medical history, which could provide deeper insights into drug prescription patterns. Furthermore, exploring hybrid modeling approaches, such as combining Random Forest with SVM, could refine decision boundaries and increase robustness in complex classification scenarios. Lastly, external validation on diverse patient datasets is essential to ensure the model's reliability across different demographics and medical conditions.

These findings suggest that Random Forest is the most effective model for drug classification, offering both high accuracy and interpretability. By refining hyperparameters, incorporating additional features, and validating the model on external datasets, its real-world applicability in medical decision-making can be significantly enhanced.

3.2. Discussion

The findings of this study highlight the varying effectiveness of machine learning models in drug classification, with Random Forest emerging as the most reliable and accurate model. The performance comparison underscores key differences between K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, each exhibiting unique strengths and limitations in terms of predictive accuracy, generalization ability, and computational complexity.

The results indicate that Random Forest achieved a perfect generalization score of 100%, demonstrating its robust predictive power and ability to handle complex medical data. In contrast, SVM exhibited strong but slightly lower accuracy (97.50% test score), while KNN lagged behind with a test score of 70%, suggesting limited generalization ability. The significant performance gap between KNN and the other two models suggests that distance-based classification methods may not be well-suited for complex medical datasets, especially those containing both categorical and continuous variables.

The inferior performance of KNN can be attributed to its reliance on distance-based classification, making it highly sensitive to variations in feature scales. Since patient attributes such as blood pressure, cholesterol levels, and sodium-to-potassium (Na/K) ratio play a crucial role in drug classification, KNN's inability to capture intricate relationships between these features likely led to its lower predictive accuracy. This limitation suggests that feature scaling and transformation techniques, such as normalization or principal component analysis (PCA), could improve KNN's performance in future studies.

The SVM model, which achieved a 97.50% test accuracy, demonstrated superior generalization compared to KNN. Its ability to effectively differentiate between

categorical and numerical features allowed it to capture meaningful patterns in the dataset. However, the computational complexity of SVM remains a challenge, especially in large-scale medical applications. The necessity for careful hyperparameter tuning—including kernel selection, regularization parameters, and margin optimization—adds another layer of difficulty when deploying SVM in real-world healthcare settings.

Random Forest outperformed both KNN and SVM due to its ensemble learning nature, which combines multiple decision trees to improve prediction accuracy and mitigate overfitting. The model's ability to capture non-linear interactions between features—such as how blood pressure, cholesterol levels, and Na/K ratios correlate—contributed to its superior generalization performance. Additionally, Random Forest provides feature importance rankings, allowing medical professionals to identify the most critical factors influencing drug classification. This interpretability makes it highly valuable in clinical decision-making.

Another advantage of Random Forest is its resilience to overfitting, which is a common issue in machine learning models. While highly complex models tend to memorize training data rather than generalizing from it, Random Forest reduces this risk by aggregating multiple decision trees, each trained on different subsets of the data. This ensemble approach enhances model robustness, making it suitable for medical applications where accurate and explainable predictions are crucial.

Despite its strong performance, Random Forest has higher computational demands than simpler models like KNN. Training and deploying Random Forest in large-scale healthcare systems may require significant processing power, which can be a limiting factor when working with extensive datasets. Future research should explore techniques such as reducing the number of estimators, optimizing tree depth, and implementing feature selection to enhance efficiency while maintaining high predictive accuracy.

The findings of this study have important implications for the use of machine learning in medical decision-making. Given its high accuracy and ability to handle complex relationships between patient attributes, Random Forest could be integrated into clinical decision support systems (CDSS) to assist healthcare providers in drug prescription. The model's ability to rank feature importance also makes it a valuable tool for identifying key patient characteristics that influence drug selection, potentially leading to more personalized treatment plans. However, before deploying machine learning models in real-world medical settings, additional considerations must be addressed:

1) External Validation: The model should be tested on larger, more diverse datasets to ensure its reliability across different patient demographics and medical conditions.

2) Feature Expansion: Additional patient-specific attributes such as lifestyle habits, genetic factors, and medical history should be incorporated to further refine drug classification accuracy.
3) Hybrid Modeling Approaches: Combining Random Forest with other machine learning techniques, such as SVM or deep learning models, could enhance classification performance by refining decision boundaries.
4) Computational Efficiency: Optimization techniques, such as reducing the number of trees in Random Forest, should be explored to balance accuracy with processing speed, making the model more feasible for real-time applications.

While this study demonstrates the effectiveness of Random Forest in drug classification, there are certain limitations that should be addressed in future research. First, the dataset size and diversity may impact model generalizability, requiring validation on real-world clinical datasets. Second, hyperparameter tuning could further enhance model performance, particularly for SVM and Random Forest, where optimal configurations may improve accuracy and efficiency. Lastly, interpretability remains a challenge for machine learning models in healthcare, and efforts should be made to integrate explainable AI (XAI) techniques to provide clearer insights into model predictions.

In summary, the Random Forest model demonstrated the highest accuracy and generalization ability, making it the most effective machine learning approach for drug classification in this study. While SVM also performed well, its computational complexity poses challenges for real-world implementation. KNN, despite its simplicity, showed limited generalization capability, highlighting the need for feature scaling and transformation techniques to improve performance. By integrating Random Forest into clinical decision-support systems, healthcare providers can enhance drug prescription accuracy and optimize patient treatment plans. Future studies should validate these findings on larger datasets, explore feature expansion, and investigate hybrid modeling techniques to further improve predictive performance in medical applications.

4. CONCLUSION

This study compared KNN, SVM, and Random Forest for drug classification based on key patient attributes. Random Forest emerged as the most effective model, achieving perfect test accuracy (100%) due to its ensemble learning approach, which enhances predictive power and reduces overfitting. SVM performed well (97.50% test accuracy) but required careful hyperparameter tuning, while KNN had the lowest accuracy (70%), struggling with mixed data types and feature scaling. These findings highlight Random Forest's potential for medical decision-making, offering high accuracy, interpretability, and resilience in complex

datasets. Future work should validate results on larger, diverse datasets, optimize hyperparameters, and explore additional patient-specific factors. Expanding AI-driven approaches in healthcare could enhance drug classification and improve treatment outcomes.

REFERENCES

[1] C. Silpa, B. Sravani, D. Vinay, C. Mounika, and K. Poorvitha, "Drug Recommendation System in Medical Emergencies using Machine Learning," in *Proc. Int. Conf. Innov. Data Commun. Technol. Appl. (ICIDCA)*, 2023, pp. 107–112, doi: 10.1109/ICIDCA56705.2023.10099607.

[2] C. Chen, "Research on Drug Classification Using Machine Learning Model," *Highlights Sci. Eng. Technol. (EMIS)*, vol. 2023, p. 350, 2024, doi: 10.54097/nfpj0845.

[3] A. Harry, "Revolutionizing Healthcare: How Machine Learning is Transforming Patient Diagnoses—A Comprehensive Review of AI's Impact on Medical Diagnosis," *BULLET: J. Multidiscip. Sci.*, vol. 2, pp. 1259–1266, 2023.

[4] S. Crisafulli, A. Fontana, L. L'Abbate, G. Vitturi, A. Cozzolino, D. Gianfrilli, M. C. De Martino, B. Amico, C. Combi, and G. Trifirò, "Machine learning-based algorithms applied to drug prescriptions and other healthcare services in the Sicilian claims database to identify acromegaly as a model for the earlier diagnosis of rare diseases," *Sci. Rep.*, vol. 14, no. 1, p. 6186, 2024, doi: 10.1038/s41598-024-56240-w.

[5] F. Aldi, I. Nozomi, and S. Soeheri, "Comparison of Drug Type Classification Performance Using KNN Algorithm," *SinkrOn*, vol. 7, no. 3, pp. 1028–1034, Jul. 2022, doi: 10.33395/sinkron.v7i3.11487.

[6] B. A. Badwan, G. Liaropoulos, E. Kyrodimos, D. Skaltsas, A. Tsirigos, and V. G. Gorgoulis, "Machine learning approaches to predict drug efficacy and toxicity in oncology," *Cell Rep. Methods*, vol. 3, no. 2, 2023, doi: 10.1016/j.crmeth.2023.100413.

[7] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, "Machine Learning in Drug Discovery: A Review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 1947–1999, Mar. 2022, doi: 10.1007/s10462-021-10058-4.

[8] H. Zhao, J. Zhong, X. Liang, C. Xie, and S. Wang, "Application of machine learning in drug side effect prediction: databases, methods, and challenges," *Front. Comput. Sci.*, vol. 19, no. 5, p. 195902, 2025, doi: 10.1007/s11704-024-31063-0.

[9] F. Aldi, I. Nozomi, and S. Soeheri, "Comparison of Drug Type Classification Performance Using KNN Algorithm," *SinkrOn*, vol. 7, no. 3, pp. 1028–1034, Jul. 2022, doi: 10.33395/sinkron.v7i3.11487.

[10] R. Hoque, M. Billah, A. Debnath, S. M. S. Hossain, and N. B. Sharif, "Heart Disease Prediction using SVM," *Int. J. Sci. Res. Arch.*, vol. 11, no. 2, pp. 412–420, Mar. 2024, doi: 10.30574/ijsra.2024.11.2.0435.

[11] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, pp. 1208–1215, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.

[12] A. Rajdhan, A. Agarwal, and M. Sai, "Heart Disease Prediction using Machine Learning," *Int. J. Eng. Res. Technol. (IJERT)*, no. 4, Apr. 2020, doi: 10.17577/IJERTV9IS040614.

[13] R. N. Ndanuko, R. Ibrahim, R. A. Hapsari, E. P. Neale, D. Raubenheimer, and K. E. Charlton, "Association between the urinary sodium to potassium ratio and blood pressure in adults: A systematic review and meta-analysis," *Adv. Nutr.*, vol. 12, no. 5, pp. 1751–1767, 2021, doi: 10.1093/advances/nmab036.

[14] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo Jr., D. W. Jones, *et al.*, "The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: The JNC 7 report," *JAMA*, vol. 289, no. 19, pp. 2560–2571, 2003.

[15] B. Lepri, J. Staiano, D. Sangokoya, E. Letouzé, and N. Oliver, "The tyranny of data? The bright and dark sides of data-driven decision-making for social good," in *Transparent Data Mining for Big and Small Data*, Springer, 2017, pp. 3–24.

[16] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python*, O'Reilly Media, Inc, 2017.

[17] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery," *J. Comput. Aided Mol. Des.*, vol. 36, no. 5, pp. 355–362, May 2022, doi: 10.1007/s10822-022-00442-9.

[18] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Overfitting, Model Tuning, and Evaluation of Prediction Performance," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer Int. Publ., 2022, pp. 109–139, doi: 10.1007/978-3-030-89010-0_4.

[19] M. Rizki, A. Hermawan, and D. Avianto, "Optimization of Hyperparameter K in K-Nearest Neighbor Using Particle Swarm Optimization," *JUITA: J. Inform.*, vol. 12, no. 1, pp. 71–79, 2024.

[20] N. Gul, M. Aamir, S. Aldahmani, and Z. Khan, "A Weighted k-Nearest Neighbours Ensemble with added Accuracy and Diversity," *IEEE Access*, vol. 10, pp. 125920–125929, Nov. 2022, doi: 10.1109/ACCESS.2022.3225682.

[21] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Inf.*, vol. 15, no. 4, 2024, doi: 10.3390/info15040235.

[22] J. Yang, Z. Wu, K. Peng, P. N. Okolo, W. Zhang, H. Zhao, and J. Sun, "Parameter selection of Gaussian kernel SVM based on local density of training set," *Inverse Probl. Sci. Eng.*, vol. 29, no. 4, pp. 536–548, 2021, doi: 10.1080/17415977.2020.1797716.

[23] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, pp. 5497–5506, Oct. 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.

[24] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[25] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylon. J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[26] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylon. J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[27] N. S. Thomas and S. Kaliraj, "An Improved and Optimized Random Forest Based Approach to Predict the Software Faults," *SN Comput. Sci.*, vol. 5, no. 5, Jun. 2024, doi: 10.1007/s42979-024-02764-x.