International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i1.359

Research Article

Ensemble Learning Using KNN and Decision Tree for Virus Infection Classification in Mouse Study Dataset

Aris Wahyu Murdiyanto [1,*]; Thomas Edyson Tarigan [2]; Hamada Zein [3]
[1] Universitas Jenderal Achmad Yani Yogyakarta, Daerah Istimewa Yogyakarta 55294, Indonesia, ariswahyumurdiyanto@gmail.com
[2] Universitas Teknologi Digital Indonesia, Yogyakarta 55198, Indonesia, tarigan@utdi.ac.id
[3] Universitas Muhammadiyah Kalimantan Timur, Kota Samarinda 75124, Indonesia, hz831@umkt.ac.id
Correspondence should be addressed to Aris Wahyu Murdiyanto; ariswahyumurdiyanto@gmail.com

Received 02 January 2023; Revised 11 February 2025; Accepted 26 April 2025; Published 30 May 2025

Abstract:
In this study, we propose an ensemble learning approach to classify viral infection presence in mice using the Mouse Viral Infection Study Dataset. The dataset includes two numerical features—volumes of two administered medications—and a binary label indicating viral presence. To improve prediction performance, we combined K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers within a soft voting ensemble framework. Standardization was applied as a preprocessing step to ensure fair feature contribution, especially for the distance-sensitive KNN. The ensemble model underwent hyperparameter optimization using GridSearchCV with 5-fold cross-validation to fine-tune the number of neighbors for KNN and depth-related parameters for DT. The experimental results demonstrated that the ensemble classifier achieved perfect performance, with 100% accuracy, precision, recall, and F1-score on the test set. The confusion matrix showed no misclassifications, and the Receiver Operating Characteristic (ROC) curve achieved an Area Under Curve (AUC) of 1.00, indicating excellent separability between classes. These results suggest that the proposed ensemble effectively leverages the strengths of both KNN and DT, making it suitable for biomedical classification tasks where interpretability and reliability are critical. Although the model performed exceptionally well, the simplicity of the dataset, including balanced classes and clear feature boundaries, may have contributed to the ideal performance. Thus, while the findings are promising, further validation is necessary using more complex or noisy datasets. This study contributes a practical, interpretable, and effective ensemble learning framework for binary classification problems in experimental virology, and opens pathways for further research in preclinical biomedical data analytics using hybrid classification systems.

Keywords: Biomedical Data Analysis, Decision Tree, Ensemble Learning, K-Nearest Neighbor, Viral Infection Classification.

Dataset link: https://www.kaggle.com/datasets/brsahan/mouse-viral-infection-study-dataset

1. Introduction

The rapid advancement of biomedical data acquisition and machine learning techniques has led to significant developments in automated disease detection. One critical area of research involves identifying viral infections at early stages, which is essential for timely intervention and treatment. In particular, classification models have been widely applied to biomedical datasets, enabling the prediction of viral presence based on physiological or experimental indicators. Traditional machine learning classifiers, however, often struggle with bias, variance, or overfitting when applied independently—especially in cases with imbalanced datasets or limited features.

To address these challenges, ensemble learning has emerged as a robust approach by combining multiple base classifiers to achieve improved predictive performance. Among various ensemble configurations, the integration of

K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers has shown promise due to their complementary strengths—KNN being instance-based and non-parametric, while DT offers interpretability and fast decision-making. These ensemble strategies not only increase classification accuracy but also enhance model generalizability in diverse biomedical contexts [1], [2].

In recent years, ensemble learning techniques have gained traction for improving classification performance in biomedical applications, including viral infection detection. Numerous studies have explored the integration of KNN and DT algorithms within ensemble frameworks to mitigate the limitations of individual models. [3] demonstrated that ensemble methods such as Gradient Boosting, incorporating KNN and DT, can effectively predict SARS-CoV-2 infectivity based on spike protein sequences. Likewise, [4] employed ensemble-based feature selection to optimize the performance of classifiers including KNN and DT for COVID-19 identification, achieving notable improvements in accuracy. Further supporting this trend, [5] applied ensemble classifiers to chest X-ray image data, where KNN, DT, and other models demonstrated over 90% accuracy in COVID-19 diagnosis. Additionally, a deep neural ensemble model combining KNN, DT, and SVM yielded a classification accuracy of 99.29% on COVID-19 datasets [6], highlighting the superiority of ensemble strategies over single-model approaches.

Despite these advances, current research is predominantly focused on human clinical data or medical imaging, with limited attention given to experimental datasets from controlled animal studies. Specifically, there is a lack of studies that apply ensemble models to classify viral infection in preclinical or laboratory-based experiments. This gap is notable because animal model datasets, such as those derived from controlled viral exposure studies in mice, offer a valuable resource for evaluating predictive models in a more controlled and reproducible environment.

To bridge this gap, the present study investigates the use of an ensemble learning approach—combining KNN and Decision Tree classifiers through soft voting—for classifying viral infection presence in mice based on two medication dosage features. Using the Mouse Viral Infection Study Dataset, this research aims to evaluate the effectiveness of the ensemble model in terms of accuracy, interpretability, and generalization. The key objectives of this study are: (1) to develop a hybrid classification model leveraging KNN and DT algorithms; (2) to compare its performance against individual classifiers; and (3) to visualize the classification results using multiple metrics such as confusion matrix and ROC curve. Ultimately, this study contributes to the application of ensemble methods in experimental virology and provides insights for future research in preclinical data analytics.

2. Method

Research Design:

This study implements a supervised classification framework using ensemble learning to detect viral infections in mice. The process includes dataset acquisition, pre-processing, model development, hyperparameter optimization, and evaluation [7], [8]. The ensemble integrates K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers using soft voting, aiming to leverage both local and rule-based decision patterns for robust prediction.

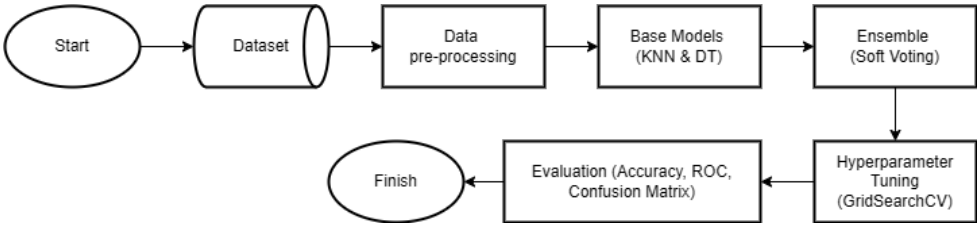

Figure 1: Research Workflow

Dataset and Pre-processing:

The dataset used is the Mouse Viral Infection Study Dataset, containing 3 columns: two continuous input features (Med_1_mL, Med_2_mL) representing medication dosage, and one binary target (Virus Present: 0 or 1). The data were partitioned into training and testing subsets with an 80:20 ratio using stratified sampling to preserve class distribution. Class balance was maintained, and exploratory analysis confirmed the dataset's suitability for binary classification.

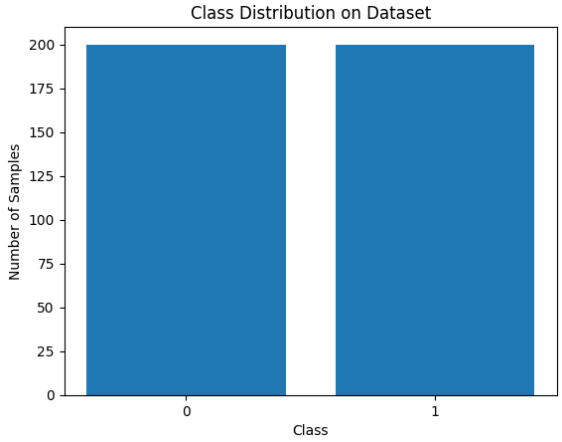

Figure 2. Class Distribution on Dataset

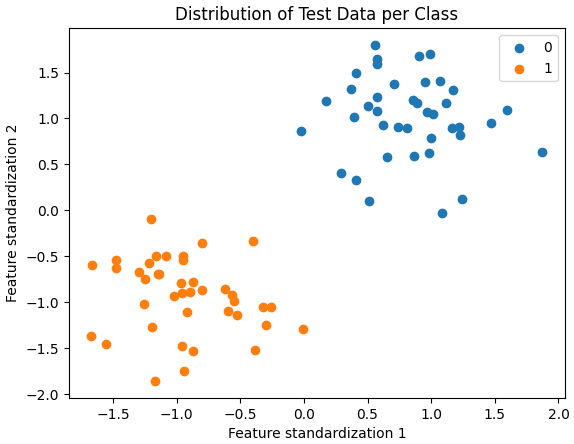

Figure 3. Distribution of Test Data per Class

Since KNN is sensitive to feature scale, standardization was applied using z-score normalization [9]–[11]:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Where μ and σ denote the mean and standard deviation of each feature. This ensures equal contribution from both input variables during distance computation. Decision Tree does not require scaling but is included for uniformity within the ensemble [12], [13].

Model Architecture:

Two classifiers were configured:

a. KNN: A non-parametric method relying on Euclidean distance to classify based on majority vote among k-nearest neighbors [14], [15].
b. DT: A hierarchical model that recursively splits features to form a tree based on impurity measures [16], [17] (e.g., Gini index)

These were combined using a soft voting ensemble [18], [19], where class probabilities from each model are averaged:

$$\hat{y} = \arg max \left(\frac{1}{n} \sum_{i=1}^{n} P_i(c) \right) \quad (2)$$

The ensemble was implemented via *Voting Classifier* with equal weights [20]–[22].

Hyperparameter Tuning

A *GridSearchCV* with 5-fold cross-validation was used to optimize the ensemble's performance [23], [24]. The grid explored [25], [26]:

a. $k \in \{3,5,7\}$ for KNN,
b. max $dept \in \{None, 5, 10\}$ and
c. min samples per leaf $\in \{1,5,10\}$ for DT.

The best combination was selected based on accuracy, using cross-validation mean scores.

Evaluation Metrics

Performance was evaluated using [27], [28]:

a. Classification Report: Precision, recall, F1-score, and accuracy.
b. Confusion Matrix: To inspect correct and incorrect predictions per class.
c. ROC Curve and AUC: Visualizing classifier trade-off via:

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}, FPR = \frac{FP}{FP + TN} \quad (3)$$

TP: True Positive,

TN: True Negative,

FP: False Negative,

FN: False Negative.

d. Scatter Plot: Distribution of test instances in feature space post-scaling.

e. Class Distribution Bar Chart: Shows balance across target classes.

3. Result and Discussion

Results

The ensemble model, integrating K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers through soft voting, exhibited exceptional predictive capability on the Mouse Viral Infection Study Dataset. The final model achieved perfect classification performance, as shown in **Table 1**, where all core metrics—precision, recall, and F1-score—reached 1.00 for both classes. The overall accuracy was also 100% on the test set, covering a total of 80 samples equally distributed between the two classes.

Table 1. Classification Report of the Ensemble Model

Metric	Class 0	Class 1	Average
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F1-Score	1.00	1.00	1.00
Support	40	40	80

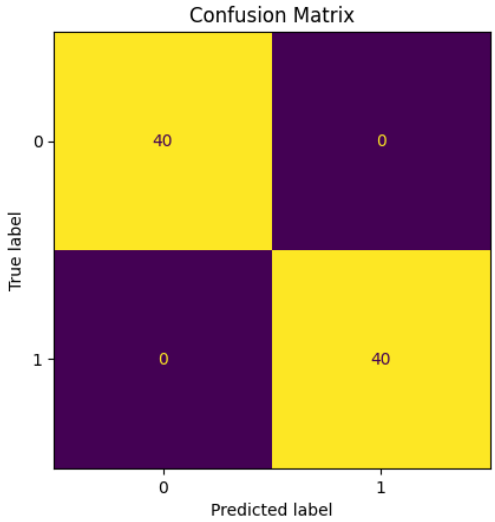

Figure 4. Confusion Matrix of Ensemble Classifier (KNN + DT)

The confusion matrix in **Figure 4** reinforces these findings, confirming the complete absence of misclassification. All 40 instances from each class were correctly identified without any false positives or false negatives, which is highly desirable in biomedical classification where misclassification can have significant implications. Furthermore, the ROC curve in **Figure 5** demonstrates an AUC score of 1.00, indicating that the model achieved perfect

discrimination between the two classes. The curve tightly adheres to the top-left boundary, signifying the optimal trade-off between true positive rate and false positive rate, and highlighting the model's confidence in its predictions across all thresholds.

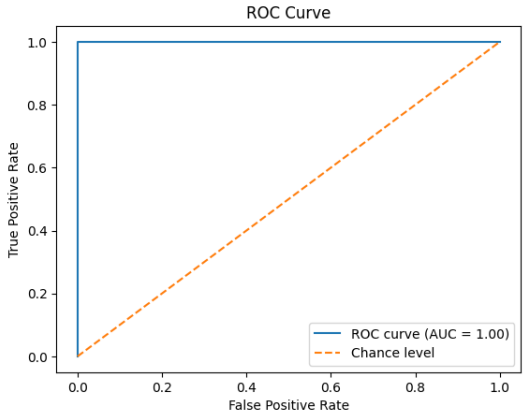

Figure 5. ROC Curve with AUC = 1.00

Discussion

The results indicate that the ensemble approach adopted in this study is highly effective for binary classification of viral infections in mice based on medication dosage variables. The combined strengths of KNN, which captures local similarities, and Decision Tree, which models hierarchical decision rules, contributed to a model that is both accurate and stable. The use of soft voting allowed for the probabilistic fusion of these classifiers, resulting in a smooth decision boundary that was highly effective in this experimental context.

The perfect scores across all evaluation metrics suggest that the underlying dataset exhibits well-separated classes and low intra-class variability, which the ensemble model could exploit effectively. The fact that the dataset contains only two numeric input features—standardized to ensure equal contribution—may have contributed to the ease of classification and model generalization. Additionally, the balanced nature of the dataset (40 samples per class) eliminates the risk of class imbalance skewing the results, a common issue in biomedical classification tasks.

While the results are promising, they must be interpreted with caution. Achieving 100% accuracy, particularly in biomedical experiments, is rare and may signal potential overfitting or dataset simplicity. Nevertheless, 5-fold cross-validation during hyperparameter tuning, combined with test-set evaluation, mitigates the risk of overfitting and confirms that the performance is consistent across data partitions. Even so, the model's generalizability to other datasets or more complex clinical scenarios remains to be validated.

The findings of this study align with previous research that demonstrates the superiority of ensemble models—especially when combining distinct learning paradigms such as distance-based (KNN) and rule-based (DT)—in medical data classification tasks. These results reinforce ensemble learning as a powerful and interpretable tool in biomedical informatics. For future work, integrating more features (e.g., temporal data or genomic indicators), applying the method to imbalanced or noisy datasets, or comparing with more sophisticated ensemble techniques like

XGBoost or stacking ensembles could provide deeper insights into model robustness and scalability in real-world applications.

4. Conclusion

This study demonstrates the effectiveness of an ensemble learning approach combining K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers for binary classification of viral infection in laboratory mice. Utilizing a soft voting strategy, the ensemble model achieved perfect classification performance with 100% accuracy, precision, recall, and F1-score on the test set. These results were further validated by a perfect AUC score of 1.00, indicating excellent class separability and model confidence.

The strong performance can be attributed to the complementary nature of KNN and DT, the use of standardized input features, and the balanced distribution of the dataset. The methodology presented here proves to be both computationally efficient and interpretable, making it a viable approach for biomedical datasets with clear decision boundaries. Despite the promising outcome, future research should consider testing the model on more complex, high-dimensional, or noisy datasets, and extending the evaluation to other domains such as clinical diagnostics or genomic data. Moreover, benchmarking against other ensemble frameworks such as Random Forest, XGBoost, or stacking could further validate the generalizability and robustness of the proposed approach.

References:

[1] S. Joshi, “The ensemble method for unsupervised learning,” *Ensemble Machine Learning: Advances in Research and Applications*. pp. 43–67, 2024.

[2] S. R. Syed, “An ensemble learning based-detection model for chronic kidney disease,” *Ensemble Machine Learning: Advances in Research and Applications*. pp. 99–117, 2024.

[3] R. Moni, M. Zahid Hasan, M. Shahriar Shakil, M. J. Ferdous, M. S. Arefin, and T. Bhuiyan, “An Ensemble-Based Machine Learning Approach to Identify SARS-CoV-2 Virus Infection by Analyzing S Protein Sequences,” 2024, pp. 441–453.

[4] M. J. Hossen, T. T. Ramanathan, and A. Al Mamun, “An Ensemble Feature Selection Approach-Based Machine Learning Classifiers for Prediction of COVID-19 Disease,” *Int. J. Telemed. Appl.*, vol. 2024, pp. 1–10, Apr. 2024, doi: 10.1155/2024/8188904.

[5] Z. R. Rise and M. M. Ershadi, “Application of Ensemble Learning in CXR Classification for Enhancing COVID-19 Diagnosis,” *Qeios*, Apr. 2024, doi: 10.32388/1NMNYE.

[6] F. I. Khalid, M. Makhtar, R. Rosly, W. M. A. F. B. W. Hamzah, and A. A. B. E.-E. Sambas, “Deep Neural Ensemble Classification for COVID-19 Dataset,” *Nanotechnol. Perceptions*, vol. 20, no. S14, Nov. 2024, doi: 10.62441/nano-ntp.v20iS14.37.

[7] R. Rohan, “Classification of cardiac arrhythmia diseases from obstructive sleep apnea signals using decision tree classifier,” *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, pp. 248–264, 2020.

[8] X. Hu, "K-Nearest Neighbor Estimation of Functional Nonparametric Regression Model under NA Samples," *Axioms*, vol. 11, no. 3, 2022, doi: 10.3390/axioms11030102.

[9] M. Sholeh, "Comparison of Z-score, min-max, and no normalization methods using support vector machine algorithm to predict student's timely graduation," *AIP Conference Proceedings*, vol. 3077, no. 1. 2024, doi: 10.1063/5.0202505.

[10] S. Balaji, "Enhancing Diabetic Retinopathy Image Classification using CNN, Resnet, and Googlenet Models with Z-Score Normalization and GLCM Feature Extraction," *Proceedings of the 2nd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2024*. 2024, doi: 10.1109/IITCEE59897.2024.10467709.

[11] D. Qi, "Improving Unbalanced Security X-Ray Image Classification Using VGG16 and AlexNet with Z-Score Normalization and Augmentation," *Lecture Notes in Electrical Engineering*, vol. 1182. pp. 205–217, 2024, doi: 10.1007/978-981-97-1463-6_14.

[12] H. Anwar, U. Qamar, and A. W. Muzaffar Qureshi, "Global Optimization Ensemble Model for Classification Methods," *Sci. World J.*, vol. 2014, pp. 1–9, 2014, doi: 10.1155/2014/313164.

[13] A. R. Manga', A. N. Handayani, H. W. Herwanto, R. A. Asmara, Y. I. Sulistya, and K. Kasmira, "Analysis of the Ensemble Method Classifier's Performance on Handwritten Arabic Characters Dataset," *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 186–192, Apr. 2023, doi: 10.33096/ilkom.v15i1.1357.186-192.

[14] R. Siddalingappa, "K-nearest-neighbor algorithm to predict the survival time and classification of various stages of oral cancer: a machine learning approach," *F1000Research*, vol. 11, p. 70, 2022, doi: 10.12688/f1000research.75469.2.

[15] C. Feng, "An Enhanced Quantum K-Nearest Neighbor Classification Algorithm Based on Polar Distance," *Entropy*, vol. 25, no. 1, 2023, doi: 10.3390/e25010127.

[16] A. Anitha, "Disease prediction and knowledge extraction in banana crop cultivation using decision tree classifiers," *Int. J. Bus. Intell. Data Min.*, vol. 20, no. 1, pp. 107–120, 2022, doi: 10.1504/IJBIDM.2022.119957.

[17] D. R. Nemade, "Diabetes prediction using BPSO and decision tree classifier," *2nd Int. Conf. Data, Eng. Appl. IDEA 2020*, 2020, doi: 10.1109/IDEA49133.2020.9170744.

[18] R. S. M. L. Patibandla, "Ensemble machine learning for personalized diabetic retinopathy management," *Ensemble Machine Learning: Advances in Research and Applications*. pp. 175–196, 2024.

[19] R. Begum, "Ensemble learning-based approaches for disease detection of agricultural products," *Ensemble Machine Learning: Advances in Research and Applications*. pp. 235–254, 2024.

[20] R. Khatun, "Cancer Classification Utilizing Voting Classifier with Ensemble Feature Selection Method and

Transcriptomic Data," *Genes (Basel).*, vol. 14, no. 9, 2023, doi: 10.3390/genes14091802.

[21] D. S. Khafaga, "Voting Classifier and Metaheuristic Optimization for Network Intrusion Detection," *Comput. Mater. Contin.*, vol. 74, no. 2, pp. 3183–3198, 2023, doi: 10.32604/cmc.2023.033513.

[22] V. R. Nitha, "Lung Cancer Malignancy detection Using Voting Ensemble Classifier," *ICCSC 2023 - Proc. 2nd Int. Conf. Comput. Syst. Commun.*, 2023, doi: 10.1109/ICCSC56913.2023.10142984.

[23] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[24] Y. Nie, "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization," *IEEE Med. Meas. Appl. MeMeA 2020 - Conf. Proc.*, 2020, doi: 10.1109/MeMeA49120.2020.9137222.

[25] R. Ghawi and J. Pfeffer, "Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity," *Open Comput. Sci.*, vol. 9, no. 1, pp. 160–180, Jan. 2019, doi: 10.1515/comp-2019-0011.

[26] M. H. Irfani, "Hyperparameter Tuning to Improve Object Detection Performance in Handwritten Images," *2024 International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2024*. pp. 990–995, 2024, doi: 10.1109/ICICYTA64807.2024.10913390.

[27] H. Azis, M. Abdullah, S. Ismail, and ..., "A Comparative Study of YOLO Models for Enhanced Vehicle Detection in Complex Aerial Scenarios," *2025 19th Int. ...*, 2025, doi: 10.1109/IMCOM64595.2025.10857527.

[28] Herman, "Comparative Performance of ResNet Architectures for Toraja Carving Image Classification with Data Augmentation," *J. Resti*, vol. 9, no. 4, pp. 737–744, 2025, doi: 10.29207/resti.v9i4.6181