Analyzing The Effect Of Feature Selection On K-Nearest Neighbor Performance In Hepatitis C Classification

Muhammmad Abduh Al Bathani, Muhammad Amin

Abstract
This study analyzes the effect of feature selection on the performance of the K-Nearest Neighbor (KNN) algorithm in classifying categories in the UCI Hepatitis C dataset, which contains 615 observations, 12 predictor features, and five target classes. Three scenarios were compared: KNN using all features, KNN with Chi-Square feature selection, and KNN with the RChiFE approach, operationalized through Tukey clipping, imputation, scaling, and Chi-Square feature selection. Evaluation used nested stratified cross-validation (outer 5-fold and inner 3-fold), with parameter selection based on macro-F1. The results show that RChiFE achieved a mean accuracy of 0.9187 (SD 0.0287) and macro-F1 of 0.5079 (SD 0.0973), compared with 0.9008 accuracy and 0.3882 macro-F1 for KNN using all features, and 0.8911 accuracy and 0.3818 macro-F1 for Chi-Square. These findings indicate improved average performance for RChiFE in this experiment, although recall for minority classes remains a concern.
***Keywords**:* Hepatitis C, K-Nearest Neighbor, Feature Selection, Chi-Square, RChiFE, Classification.

Muhammad Amin
Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: mhdamin@dosen.pancabudi.ac.id

Muhammmad Abduh Al Bathani
e-mail: m.abduh40@gmail.com
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Hepatitis C virus (HCV) infection is a health concern that requires accurate detection and assessment of liver conditions. Laboratory and demographic data can be used to develop machine-learning-based classification approaches. The HCV dataset available in the UCI Machine Learning Repository contains data from blood donors and patients with Hepatitis C, including disease categories such as hepatitis, fibrosis, and cirrhosis [1].

The K-Nearest Neighbor (KNN) algorithm is a classification method based on distance between samples. Its performance is affected by feature scales, data distribution, missing values, outliers, and the number of neighbors (K). In clinical data containing laboratory variables with different units and ranges, preprocessing and feature selection are important steps in model development.

Feature selection can reduce input dimensionality while retaining attributes that are statistically associated with the target class. Chi-Square is a filter method that measures the association between features and classes. Studies on feature selection for Hepatitis C classification indicate that attribute selection is a relevant stage to evaluate alongside classifier performance [2]. Other studies using the UCI HCV dataset also highlight the importance of preprocessing, class imbalance handling, and feature selection [5].

This study compares KNN without feature selection, KNN with standard Chi-Square selection, and KNN with the RChiFE approach. In this study, RChiFE is operationally defined as a process that limits numerical outliers using Tukey fences calculated from training data, imputes missing values, encodes categorical variables, applies Min-Max scaling, and ranks features using the Chi-Square statistic. This definition is stated explicitly to support reproducibility and to avoid presenting RChiFE as a universally standardized algorithm.

The contribution of this study is a controlled evaluation of the effect of feature selection on KNN performance for five-category HCV classification. Nested cross-validation is used so that feature and parameter selection are performed only on the training portion of each iteration. This approach is intended to reduce evaluation bias caused by information leakage.

Literature Review

1. UCI Hepatitis C Dataset

The UCI HCV Data (ID 571) consists of 615 observations and 12 predictor features. The features include age, sex, and the laboratory indicators ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT. The Category variable has five labels: 0=Blood Donor, 0s=Suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, and 3=Cirrhosis. The dataset contains missing values and is licensed under CC BY 4.0 [1]. The patient identification column was not used as a predictor.

2. K-Nearest Neighbor

KNN assigns a sample to the majority class among its nearest neighbors. This study uses Euclidean distance and distance-based weighting. The candidate values of K are 3, 5, 7, 9, and 11. Scaling is necessary because features with larger numerical ranges may otherwise dominate distance calculations.

3. Chi-Square and RChiFE Feature Selection

The Chi-Square test is used to assess the dependence between non-negative features and class labels. Because the Chi-Square statistic in the scikit-learn implementation requires non-negative inputs, numerical features are mapped to a non-negative range using Min-Max scaling. In this study, RChiFE additionally limits extreme values using Tukey fences calculated exclusively from training data. All steps are executed within the cross-validation pipeline so that imputation statistics, outlier thresholds, scaling parameters, and feature scores are not learned from the test fold.

4. Classification Evaluation

Accuracy measures the proportion of correct predictions. Precision, recall, and F1-score are calculated using macro averaging so that each class receives equal weight. Balanced

accuracy is used to reduce dependence on the majority class. The confusion matrix is used to examine misclassification patterns between classes.

Research Methodology

1. Research Stages

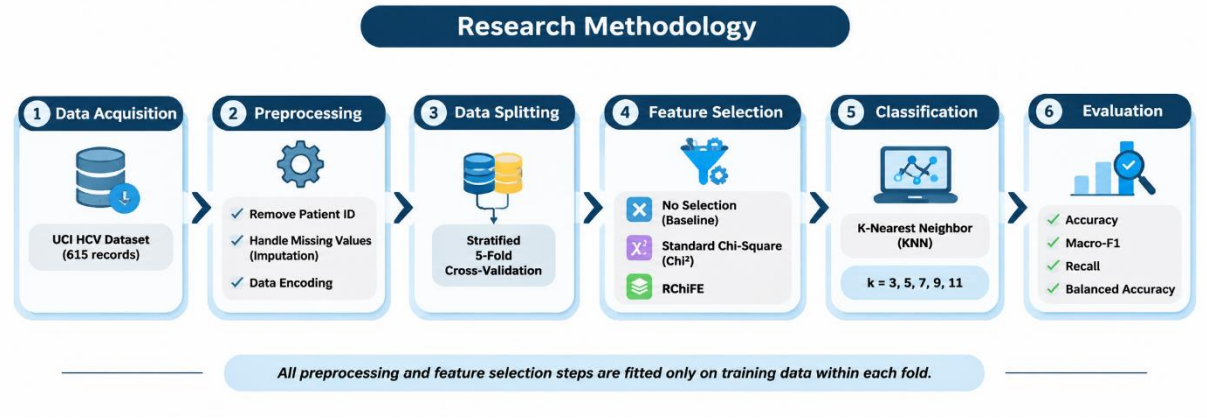


Figure 1. Research Stages

The research consists of six stages: data acquisition, preprocessing, data splitting, feature selection, KNN classification, and model evaluation. The UCI Hepatitis C dataset contains 615 records. Three feature-selection scenarios are compared: no selection, Chi-Square, and RChiFE. Model performance is evaluated using stratified 5-fold cross-validation based on accuracy, macro-F1, recall, and balanced accuracy.

2. Data Source and Characteristics

Data were obtained from the UCI Machine Learning Repository, HCV Data dataset ID 571, DOI 10.24432/C5D612. The number of observations and features follows the repository metadata. The classification target is Category. The ID column was removed from the predictor matrix. All five categories were retained to preserve the category information provided in the dataset.

The raw category distribution shows that the Blood Donor class dominates the dataset. This imbalance should be considered because accuracy alone may conceal poor performance on minority classes.

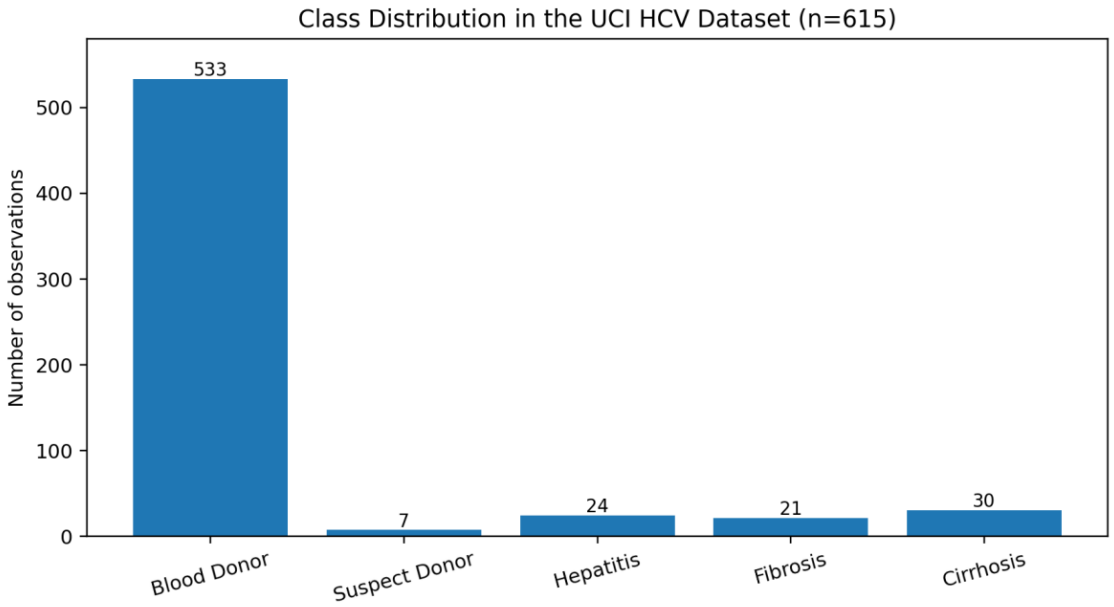


Figure 2. Distribution Of The Five Categories In The UCI HCV Dataset

3. Preprocessing and Experimental Scenarios

Numerical features were imputed using the median, while categorical features were imputed using the mode; the Sex variable was then transformed through one-hot encoding. All features were mapped to a non-negative scale. In the RChiFE scenario, numerical features were first clipped to the Tukey fences Q1−1.5×IQR and Q3+1.5×IQR. Chi-Square was then used to select the highest-scoring features. The candidate numbers of selected features were 4, 6, 8, 10, and 12.

Table 1. Model Comparison Scenarios

Scenario	Preprocessing	Feature selection	Classifier
A	Imputasi, encoding, Min-Max	Not	KNN
B	Imputasi, encoding, Min-Max	Chi-Square	KNN
C	Clipping Tukey, imputasi, encoding, Min-Max	RChiFE (operasional)	KNN

4. Validation and Parameters

Evaluation used outer Stratified 5-Fold Cross-Validation with a fixed random state (random_state=42). In each outer fold, K and the number of features were selected through inner Stratified 3-Fold Cross-Validation based on macro-F1. The best model was then evaluated on the test fold, which had not been used during parameter selection. This design constitutes nested cross-validation.

5. Reproducibility

The accompanying Python script downloads the dataset using the ucimlrepo package with dataset ID 571, builds the pipelines, runs cross-validation, and saves fold-level results, metric summaries, feature-selection frequencies, confusion matrices, and performance comparison plots. The software-library versions and execution environment should be recorded when the experiment is run.

Results

1. Dataset Characteristics

The UCI HCV dataset contains 615 records across five target categories. Based on the uploaded dataset, Blood Donor accounts for 533 observations, Suspect Blood Donor for 7, Hepatitis for 24, Fibrosis for 21, and Cirrhosis for 30. The dominance of the Blood Donor class indicates substantial class imbalance; therefore, evaluation should not rely on accuracy alone. This distribution describes the dataset and is not a model prediction result.

Table 2. Class Distribution In The UCI HCV Dataset

Category	Count	Proportion
Blood Donor	533	86.67%
Suspect Donor	7	1.14%
Hepatitis	24	3.90%
Fibrosis	21	3.41%
Cirrhosis	30	4.88%

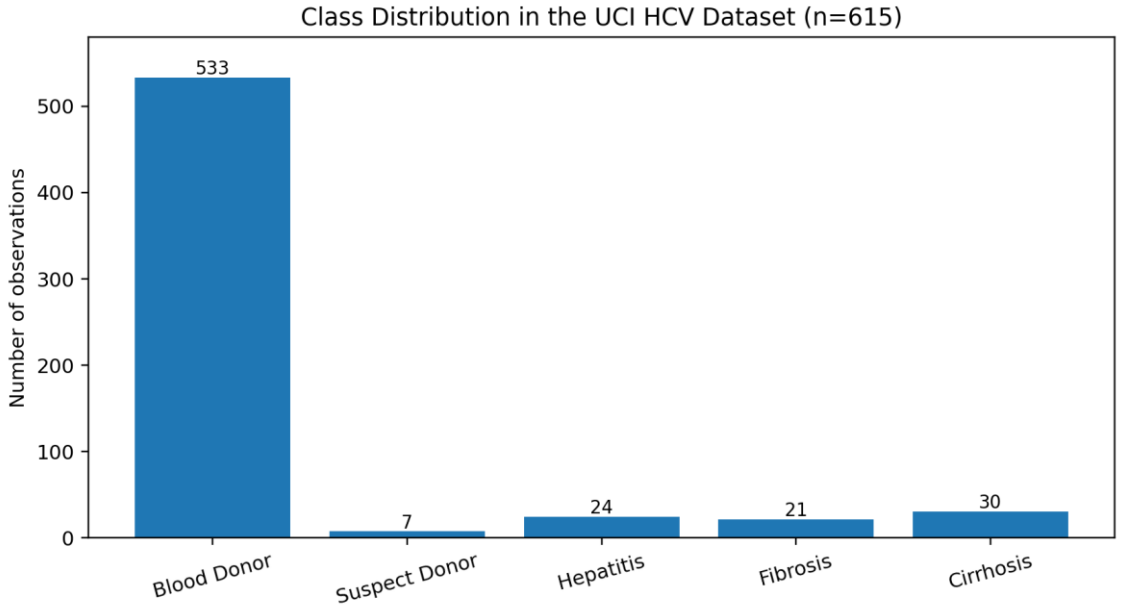


Figure 3. Category distribution in the UCI HCV dataset

2. Missing Value Examination

The dataset contains missing values in ALB (1), ALP (18), ALT (1), CHOL (10), and PROT (1), amounting to 31 missing values across five variables. Missing values must be handled using statistics learned only from the training data in each validation fold.

Table 3. Summary Of Missing Values In The Reported Variables

Variable	Number of missing values
ALB	1
ALP	18
ALT	1
CHOL	10
PROT	1

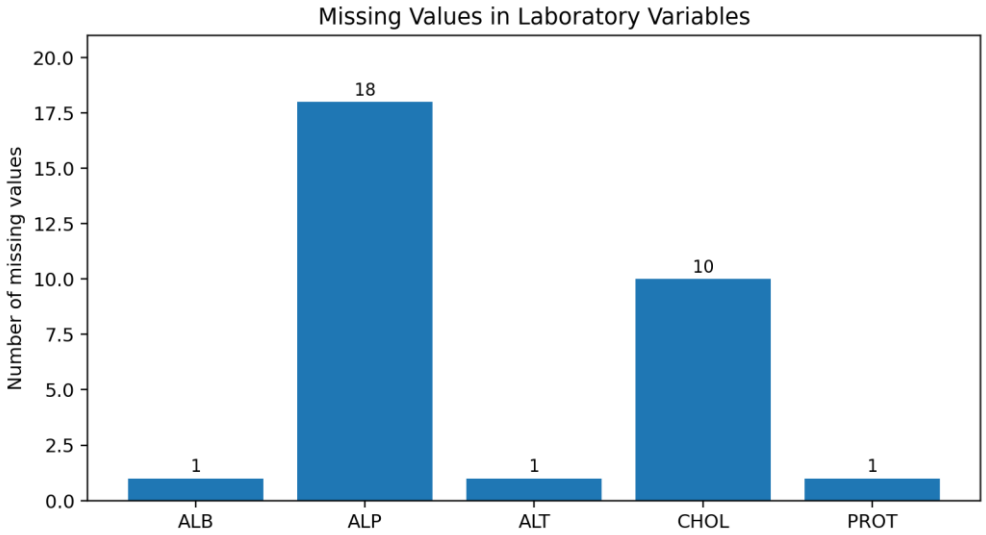


Figure 4. Number Of Missing Values In Laboratory Variables

3. Classification Experiment Results

The experiment was successfully run on the uploaded hcvdat0.csv file. All scenarios were evaluated using outer Stratified 5-Fold Cross-Validation. In each fold, the number of selected

features and K value were determined through inner Stratified 3-Fold Cross-Validation based on macro-F1. Table 5 reports the mean and standard deviation of the outer-fold metrics. These values are measurements on the outer test folds, not training scores.

As shown in Table 5, RChiFE achieved the highest mean accuracy and macro-F1 among the three evaluated scenarios. RChiFE accuracy was 0.9187, which is 0.0179 higher than KNN using all features (0.9008) and 0.0276 higher than Chi-Square (0.8911). The difference in macro-F1 was more pronounced: RChiFE achieved 0.5079, compared with 0.3882 for KNN using all features and 0.3818 for Chi-Square. This suggests that outlier handling, as operationalized in RChiFE, improved aggregate performance in this experiment. However, the standard deviations and imbalanced class distribution require cautious interpretation; the model should not be considered ready for clinical decision-making. In absolute terms, RChiFE accuracy was 1.79 percentage points higher than KNN using all features, while macro-F1 was 11.97 percentage points higher. Nevertheless, the confusion matrix shows that the recall for the Suspect Blood Donor class was zero because none of its samples was correctly predicted. Thus, improved aggregate scores do not mean that the model performs evenly across all categories.

Table 4. Summary Of Outer Cross-Validation Metrics

Model	Accuracy	Macro precision	Macro recall	Macro F1	Balanced accuracy
KNN with all features	0.9008 ± 0.0156	0.5046 ± 0.0986	0.3570 ± 0.0596	0.3882 ± 0.0608	0.3570 ± 0.0596
KNN + Chi-Square	0.8911 ± 0.0178	0.4145 ± 0.0402	0.3677 ± 0.0638	0.3818 ± 0.0574	0.3677 ± 0.0638
KNN + RChiFE	0.9187 ± 0.0287	0.6053 ± 0.1551	0.4752 ± 0.0885	0.5079 ± 0.0973	0.4752 ± 0.0885

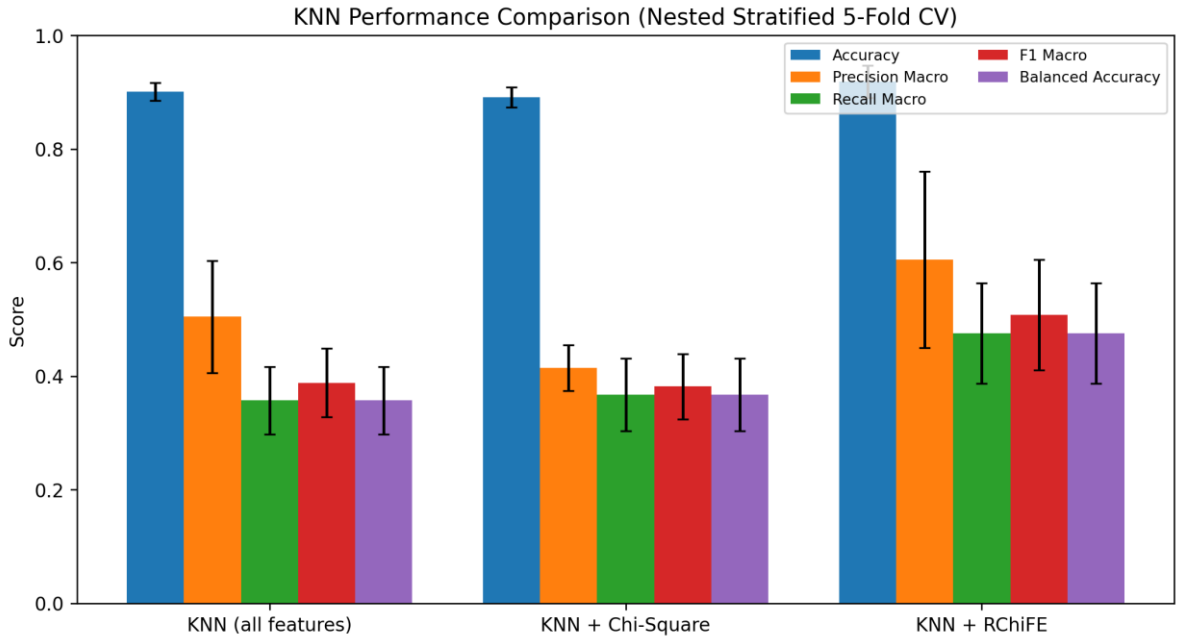


Figure 5. Comparison Of Mean Model Metrics; Error Bars Indicate Standard Deviations Across Outer Folds

The best parameter settings selected in each outer fold show that KNN using all features selected K=3 in four folds and K=5 in one fold. Chi-Square selected combinations of 4–12 features and K=3 or 5, whereas RChiFE selected 4–12 features and K=5 or 7. These parameters were selected exclusively through inner cross-validation on the training data of each outer fold.

Table 5. Parameters Selected In Each Outer Fold

Model	Fold	Number of features	K
KNN with all features	1	12	3
KNN with all features	2	12	3
KNN with all features	3	12	3
KNN with all features	4	12	5
KNN with all features	5	12	3
KNN + Chi-Square	1	12	3
KNN + Chi-Square	2	4	3
KNN + Chi-Square	3	4	5
KNN + Chi-Square	4	4	3
KNN + Chi-Square	5	6	3
KNN + RChiFE	1	4	7
KNN + RChiFE	2	12	5
KNN + RChiFE	3	6	7
KNN + RChiFE	4	12	5
KNN + RChiFE	5	12	7

4. Confusion Matrix Analysis and Feature Stability

The out-of-fold confusion matrix for RChiFE shows that most Blood Donor cases were correctly identified (529 of 533 observations). For the Cirrhosis class, 22 of 30 observations were correctly classified. In contrast, none of the 7 Suspect Blood Donor observations was correctly identified as belonging to its own class. This confirms that, although macro-F1 improved, performance on the smallest class remains limited.

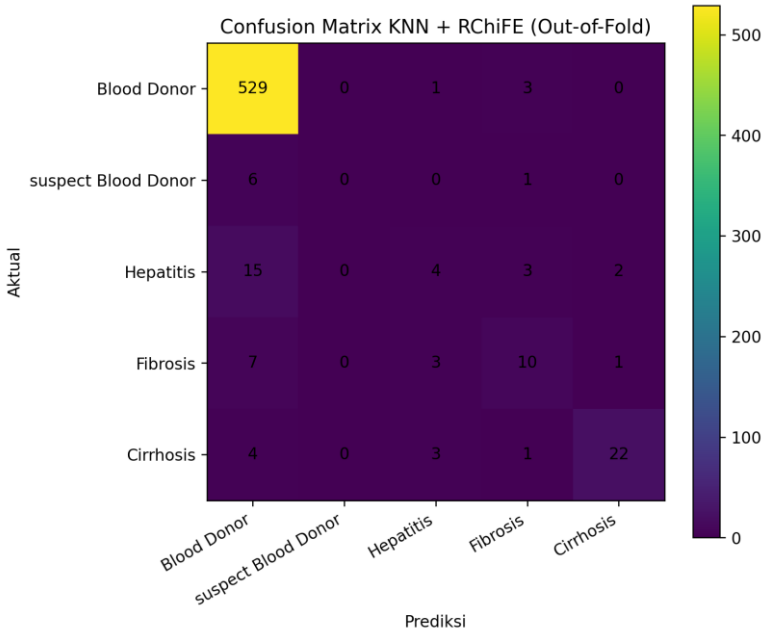


Figure 6. Out-Of-Fold Confusion Matrix For The KNN + Rchife Model

Discussion

The results indicate that RChiFE improves KNN performance, achieving the highest accuracy (91.87%) and macro-F1 score (50.79%) compared with standard KNN and Chi-Square selection. This suggests that combining Tukey-based outlier handling and feature

selection can enhance classification performance. However, the model still struggles to identify minority classes, particularly Suspect Blood Donor, indicating that class imbalance remains a major challenge. Therefore, further improvements are needed to achieve more balanced classification performance across all categories.

Limitations

This study is limited by the imbalanced distribution of the UCI Hepatitis C dataset, particularly the small number of Suspect Blood Donor records. The experiments also rely on a single dataset and do not include external validation. Although RChiFE improves overall classification performance, its ability to identify minority classes remains limited. Future research should consider class-balancing techniques and additional datasets to improve model generalization and classification reliability.

Conclusion

This study compared KNN using all features, KNN with Chi-Square feature selection, and KNN with RChiFE on the UCI HCV dataset. Based on nested stratified cross-validation, RChiFE achieved a mean accuracy of 0.9187 and macro-F1 of 0.5079, higher than the two comparison scenarios. These results support a positive effect of the operationalized RChiFE approach on aggregate KNN performance in this experiment. However, the confusion matrix shows that the Suspect Blood Donor class was not correctly identified, indicating that class imbalance remains an important limitation. Future work should examine minority-class handling strategies, external validation, and result stability on independent datasets.

References

[1] R. Lichtinghagen, F. Klawonn, and G. Hoffmann, "HCV data," UCI Machine Learning Repository, 2020, doi: 10.24432/C5D612.

[2] "Selected feature selection methods for classifying patients with Hepatitis C," *Procedia Computer Science*, vol. 225, pp. 3710–3717, 2023, doi: 10.1016/j.procs.2023.10.366.

[3] "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intelligent Medicine*, vol. 2, no. 4, pp. 193–198, 2022, doi: 10.1016/j.imed.2021.12.003.

[4] "Enlightened prognosis: Hepatitis prediction with an explainable machine learning approach," *PLOS ONE*, 2025, doi: 10.1371/journal.pone.0319078.

[5] "Prediction of hepatitis-C virus using statistical learning models," *Discover Public Health*, vol. 22, Art. no. 306, 2025, doi: 10.1186/s12982-025-00654-y.

[6] "IHCP: Interpretable hepatitis C prediction system based on black-box machine learning models," *Scientific Reports*, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10481489/.

[7] "A cross dataset meta-model for hepatitis C detection using multi-dimensional pre-clustering," *Scientific Reports*, 2025, doi: 10.1038/s41598-025-91298-0.

[8] Yusfila, "Klasifikasi penyakit Hepatitis C dengan menggunakan K-Nearest Neighbor," *Sains Data: Jurnal Studi Matematika dan Teknologi*, 2025. [Online]. Available: https://pub.nuris.ac.id/sainsdata/article/view/205.

[9] P. H. Putra and M. S. Novelan, "Perancangan aplikasi penentuan kualitas sayuran berdasarkan warna menggunakan data mining," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 103–109. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4156.

[10] Nuranisah, "Analisis menggunakan Random Forest dengan Gini Index algoritma pada data," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 19–24. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4165.

[11] A. P. Sidik, Hermansyah, and D. Aulia, "Rancangan sistem pengelompokkan buah berdasarkan kadar vitamin menggunakan metode K-Medoids," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2023. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4535.

[12] M. E. Dalimunthe, Z. Lubis, E. Sutejo, and D. P. Sari, "Analisis solar cell sebagai sumber energi alternatif pada Gedung I Universitas Pembangunan Panca Budi," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2023. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/download/4506/4090.

[13] M. D. B. Barus, Mustafa, and F. S. Thahirah, "Single eksponensial smoothing: Analisis forecasting dalam perencanaan produksi," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 199–212. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4150.

[14] M. Y. Insan, S. S. Batubara, and N. Wulandari, "Analisis faktor-faktor yang mempengaruhi keterikatan kerja perawat di Medan pada era COVID-19," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 34–43. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4168.

[15] E. S. Matondang and Juniarti, "Efektivitas penggunaan anggaran belanja langsung terhadap anggaran berbasis kinerja," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 44–52. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4169.

[16] D. A. D. Nasution, G. Supraja, and A. F. Damanik, "Faktor-faktor yang mempengaruhi kualitas laporan keuangan pada Pusat Penelitian Kelapa Sawit Medan," in *Proc. Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2022, pp. 255–265. [Online]. Available: https://jurnal.pancabudi.ac.id/index.php/scenario/article/view/4201.

[17] M. Ahmad et al., "Diagnosing the stage of Hepatitis C using machine learning," *Journal of Healthcare Engineering*, vol. 2021, Art. no. 8062410, 2021, doi: 10.1155/2021/8062410.

[18] H. Chen et al., "A survey of k nearest neighbor algorithms for solving the class imbalanced problem," *Wireless Communications and Mobile Computing*, vol. 2021, Art. no. 5520990, 2021, doi: 10.1155/2021/5520990.

[19] C. Villa-Blanco, C. Bielza, and P. Larrañaga, "Feature subset selection for data and feature streams: A review," *Artificial Intelligence Review*, vol. 56, pp. 1011–1062, 2023, doi: 10.1007/s10462-023-10546-9.

[20] A. Moslemi, "A tutorial-based survey on feature selection: Recent advancements on feature selection," *Engineering Applications of Artificial Intelligence*, vol. 126, Art. no. 107136, 2023, doi: 10.1016/j.engappai.2023.107136.

[21] "Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges," *Swarm and Evolutionary Computation*, vol. 90, Art. no. 101661, 2024, doi: 10.1016/j.swevo.2024.101661.

[22] P. Khatun, S. Umam, R. B. Razzak et al., "A study on the effectiveness of machine learning models for hepatitis prediction," *Scientific Reports*, vol. 15, Art. no. 30659, 2025, doi: 10.1038/s41598-025-07104-4.

[23] M. Tusher et al., "Comparative investigation of bagging enhanced machine learning for early detection of HCV infections using class imbalance technique with feature selection," *PLOS ONE*, vol. 20, no. 6, Art. no. e0326488, 2025, doi: 10.1371/journal.pone.0326488.

[24] F. Furizal, A. Ma'arif, and D. Rifaldi, "Application of machine learning in healthcare and medicine: A review," *Journal of Robotics and Control*, vol. 4, no. 5, pp. 621–631, 2023, doi: 10.18196/jrc.v4i5.19640.