**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

# The implementation of the Random Forest Algorithm with Resampling and Without Resampling on the Hepatitis C Disease Dataset.

**I Gede Hendrayana[1*], Ni Putu Dita Ariani Sukma Dewi[2], Jiyestha Aji Dharma Aryasa[3], I Made Ade Prayoga[4], Rizki Anom Raharjo[5)]**

[1,2,3,4,5)] Institut Teknologi dan Kesehatan Bali, Indonesia
[1)]igedehendrayana79@gmail.com, [2)]ditaariani@itekes-bali.ac.id, [3)]jiyestha.aji@gmail.com,
[4)]adeprayogaa1996@gmail.com @gmail.com, [5)] rizki.anomraharjo@gmail.com@email.com

## ABSTRACT

This study evaluates the performance of Random Forest models for Hepatitis C classification using a dataset from Kaggle, focusing on addressing class imbalance through resampling techniques. We compare three approaches: baseline Random Forest without resampling, Random Forest with SMOTE+ENN (Synthetic Minority Oversampling Technique + Edited Nearest Neighbors), and Random Forest with SMOTE+OSS (Synthetic Minority Oversampling Technique + One-Sided Selection). Results show that the baseline model achieved high accuracy (0.9837) but failed to detect minority classes (e.g., suspect Blood Donor recall=0.00). SMOTE+ENN significantly improved performance, achieving perfect classification (precision=1.00, recall=1.00) for Hepatitis, Fibrosis, and Cirrhosis, while maintaining high accuracy (0.9919) and ROC AUC (0.9999). In contrast, SMOTE+OSS showed limitations in detecting Hepatitis (recall=0.00) and yielded lower precision for Fibrosis (0.44), indicating its undersampling approach may be too aggressive. The study highlights SMOTE+ENN as the most effective method for balancing class distribution and enhancing model robustness in medical diagnostics. These findings underscore the importance of selecting appropriate resampling techniques to improve minority class detection in imbalanced datasets, with implications for developing reliable AI-based diagnostic tools for Hepatitis C.

**Keywords:** Hepatitis C, Random Forest, SMOTE, ENN, OSS

## INTRODUCTION

Hepatitis is categorized as an acute disease that can affect all age groups, from children to adults. Most acute Hepatitis infections present as mild illnesses or may remain undetected. However, in certain cases, they can progress to severe complications with fatal outcomes. In 2019, approximately 78,000 deaths were recorded globally due to complications arising from acute viral Hepatitis A-E infections (Elliott dkk., 2024; Pourkarim, 2022). In Indonesia, an estimated 20 million individuals suffer from Hepatitis, predominantly Hepatitis B. The CDA Foundation documented 51,100 annual deaths attributed to Hepatitis B and 5,942 annual deaths related to Hepatitis C in 2016. According to BPJS Kesehatan data, 2,159 deaths from cirrhosis and liver cancer - conditions commonly associated with advanced Hepatitis B were recorded in 2022 (Diniarti & Romli, 2024). Hepatitis C represents the most severe Hepatitis type. WHO estimates that 2-2.5% of the global population carries persistent HCV infection. The Hepatitis C virus belongs to the Flaviviridae family (Charisma dkk., 2022; Zhang dkk., 2025). The Hepatitis C virus is mostly transmitted through contaminated blood and can also be through sexual intercourse, but it is still rare. Hepatitis C disease is important to be detected early in order to identify the disease before symptoms appear because we can reduce the transmission rate of people with the disease to death. So identifying aims to detect the possibility of disease (Mehraeen dkk., 2024; Schwarz dkk., 2022). One of the steps that can be taken to identify the type of Hepatitis C disease is computerized using data mining techniques.

Data mining has a role in collecting, refining, handling and extracting information into valuable information (Chen dkk., 2023). Data mining is a computational approach that utilizes a variety of analytical techniques to extract hidden patterns, trends, and causal relationships in large-scale datasets, with the primary goal of generating new knowledge of strategic value (Hakim dkk., 2024). In the context of health, the integration of data mining techniques with machine learning algorithms, especially Random Forest, has proven effective for building disease prediction systems, including in the identification of Hepatitis C patients based on clinical parameters. This algorithm excels because of three main characteristics, namely the ability to generate models with minimal error rates through the ensemble learning mechanism, producing good calculation accuracy, and resistance to missing values which are

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

usually common problems in medical data (Liu dkk., 2024; Sari dkk., 2023). The implementation of Random Forest is inseparable from crucial shortcomings, namely accuracy fluctuations influenced by the composition of training data, the need for intensive computing resources, especially for high-dimensional datasets, and the reliance on precise hyperparameter tuning processes to achieve optimal performance (Iman & Wijayanto, 2021). This deficiency is an important consideration in the design of a Random Forest-based prediction system for clinical cases such as Hepatitis C. The dataset used in this study came from the Kaggle.com website with a total of 615 patients who are prone to dataset imbalances. The model tends to be biased towards the majority class so that it fails to detect Hepatitis C patients, and causes overfitting and evaluation of the model that only relies on accuracy (Wang dkk., 2025). Improve the quality of the dataset by removing ambiguous or unclear majority class data, resulting in better model training data (Ahmed dkk., 2025; Chowdhury dkk., 2024) while the SMOTE (Synthetic Minority Oversampling Technique) up sampling method is to increase synthetic samples for minority classes so that the number is balanced with the majority class (Chowdhury dkk., 2024; Kurnianingsih dkk., 2025). The use of ENN (Edited Nearest Neighbors) and OSS (One-Sided Selection) undersampling methods aims to balance the class by removing the sample from the majority class that is ambiguous or noisy and maintaining an informative sample that supports the formation of a robust model (Ahmed dkk., 2025; Mooijman dkk., 2023). This study aims to conduct a comparative analysis of the performance of the Random Forest algorithm in predicting Hepatitis C disease by considering two different data processing scenarios, namely by comparing the performance of the application of the hybrid resampling techniques SMOTE+ENN and SMOTE+OSS. The main focus of the study was to evaluate the influence of dataset balancing techniques on prediction accuracy, taking into account various evaluation metrics such as Precision, Recall, F1-score and AUC-ROC.

The results of this study are expected to provide a comparison of the combined performance of SMOTE+ENN and SMOTE+OSS in improving the quality of predictions in unbalanced data, significant comparisons between models with and without data balancing and recommendations for the implementation of the Random Forest algorithm in the Hepatitis C disease prediction system in the clinical environment. In addition, the results of this study can contribute to the development of more accurate AI-based diagnostic systems, especially for cases of diseases with uneven data distribution such as Hepatitis C.

## LITERATURE REVIEW

In the research of Mohammad Mihrab Chowdhury, et al. 2024 aim to overcome the challenge of diabetes classification in an unbalanced dataset using data augmentation techniques and machine learning algorithms. The dataset used is BRFSS 2021, with the proportion of diabetes cases (minority) only 18% compared to non-diabetic (majority) 82%. Undersampling (ENN) techniques are most effective for improving the detection of diabetes cases, while hybrid sampling provides a balanced alternative between precision and recall (Chowdhury dkk., 2024). Research by Ananda Sutradhar, et al. 2024 develops an accurate and reliable diagnostic system for thyroid disease using hybrid machine learning (ML) and Explainable AI (XAI) techniques. The datasets used were from Kaggle and UCI, with significant class imbalances (for example, only 231 cases of thyroid positive out of 3772 samples in the Kaggle dataset). The combination of SMOTE-ENN and feature selection (IGS) effectively overcomes data imbalance and dimension reduction and the use of the RDKST hybrid model excels in accuracy and stability, suitable for complex datasets (Sutradhar dkk., 2024). Research by Kwabena Ebo Bennin, et al. 2022 The use of OSS (One-Sided Selection) is an ideal choice for CPDP when precision (reducing false alarms) is more important than recall, especially after the implementation of NN Filter (Bennin dkk., 2022).

The research of Canero, et al. 2024 evaluated a Random Forest (RF)-based feature selection method to predict the content of soil organic matter (SOM), clay, and carbonates using soil spectroscopy data. The two feature selection methods tested were Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS), which were compared with conventional methods such as Partial Least Squares Regression (PLSR) and RF without feature selection (Canero dkk., 2024). The research of Shunqi Nie, et al. 2024 used a Random Forest (RF) model to predict the spatial distribution of six heavy metals (Cr, Cd, Pb, As, Hg, and Ni) in the soil of China's east coast region. Heavy metals data were collected from 748 sample points, with environmental variables such as rainfall, population density, elevation, and soil organic matter used as predictors. The results showed that the RF model performed well in predicting the distribution of heavy metals, with $R^2$ values ranging from 0.482 to 0.627 for validation data (Nie dkk., 2024). Research by Kellen Sumwiza, et al. 2023 developed a prediction model for cardiovascular disease (CVD) using a Random Forest (RF) algorithm that was enhanced by feature selection techniques. A dataset from Kaggle with 1025 data and 14 attributes was processed to eliminate outliers and irrelevant features, resulting in 769 data with 13 features. The RF model is compared with K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression (LR). The results showed that the RF model achieved the highest accuracy, which was 99%, after the

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

application of feature selection techniques, compared to KNN (95%), SVM (85%), and LR (87%) (Sumwiza dkk., 2023).

## METHOD

In this study, using the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is a stage of systematic framework preparation used in data mining management research which is the most popular framework because of its flexibility that can be applied in various industries which includes 6 stages (Schröer dkk., 2021). CRISP-DM demonstrates significant advantages over the Knowledge Discovery in Databases (KDD) approach, particularly in its more systematic and application-oriented process structure. This framework divides the data mining process into six integrated phases, ranging from business understanding to deployment, thereby enabling a stronger alignment between analytical activities and the strategic objectives of an organization. Unlike KDD, which primarily emphasizes technical aspects of pattern discovery, CRISP-DM places business context understanding at the foundation of its process, making it more relevant for data-driven decision-making. The iterative nature of CRISP-DM also offers high flexibility in adapting to project dynamics, in contrast to the linear progression characteristic of KDD. Furthermore, CRISP-DM's widespread adoption across industries as a de facto standard in data mining practice supported by comprehensive documentation and its ability to facilitate multidisciplinary collaboration underscores its superior practical applicability compared to KDD, which remains more prevalent within academic domains (Baskara dkk., 2025; Shimaoka dkk., 2024).

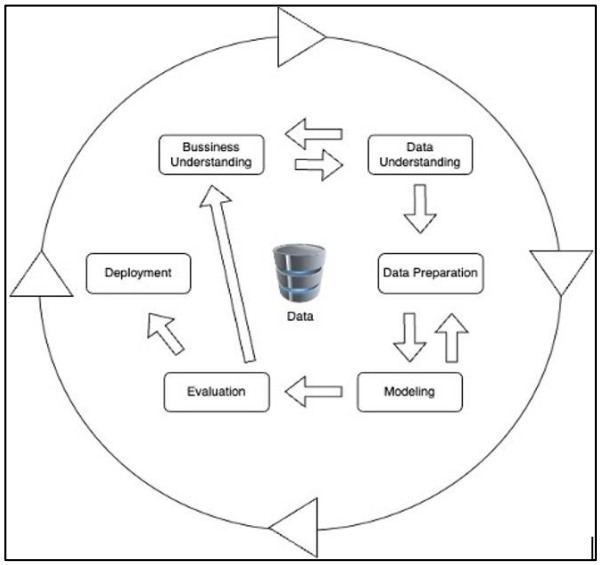



Fig. 1 Stages of CRISP-DM

1. Bussiness Understanding

In the business understanding stage, this study adopts a systematic methodology that includes the identification of clinical problems related to hepatitis C diagnosis, a comprehensive literature review on the application of machine learning in the field of hepatology, and an in-depth exploratory analysis of the characteristics of the hepatitis C dataset. dealing with data imbalances and without a resampling process aimed at building predictive models that can accurately distinguish between patients infected with hepatitis C and healthy individuals. The technical implementation was carried out using Visual Studio Code as an integrated development environment (IDE), with the aim of conducting a comparative evaluation of the effectiveness of resampling techniques in improving classification performance. The results of this study are expected to not only provide technical solutions to the problem of medical data imbalances, but also to generate inputs in the clinical field that can support the diagnostic decision-making process, especially in terms of the identification of hepatitis C positive cases and the development of evidence-based medicine-based clinical decision support systems.

2. Data Understanding

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

The data collection process in this study was carried out through a systematic exploration of various secondary data sources that are relevant to the research topic. This stage includes identification and critical evaluation of potential data sources, selection of datasets based on criteria of completeness, relevance, and credibility of sources and initial analysis to identify characteristics and potential problems in the dataset. As the primary data source, this study utilizes public datasets obtained from the Kaggle (www.kaggle.com) platform, a leading science data repository that provides a wide range of datasets for research purposes. The dataset used consisted of 615 patient medical records with 13 clinical attributes that included demographic parameters, laboratory test results and clinical indicators related to the patient's health condition. Before further analysis, this dataset goes through an initial validation process to ensure the quality of the data, including checking for potential issues such as missing values, class imbalances, or outliers that may affect the results of the analysis. The following is an explanation of the column attributes in Table 1.

Table 1. Attributes of the Hepatitis C Column

| Column Attributes | Attribute Description |
|---|---|
| Category | Target test features |
| Age | Patient's age |
| Sex | Patient's gender |
| ALB (albumin) | Amount of albumin |
| ALP (*alkaline phosphatase*) | Amount of *alkaline phosphatase* |
| ALT (*alanine transaminase*) | Amount of *alanine transaminase* |
| AST (*aspartate aminotransferase*) | Amount of *aspartate aminotransferase* |
| BIL (*bilirubin*) | Amount of *bilirubin* |
| CHE (*cholinesterase*) | Amount of *cholinesterase* |
| CHOL (*cholesterol*) | Amount of *cholesterol* |
| CREA (*creatine*) | Amount of *creatine* |
| GGT (*gamma-glutamyl transferase*) | Amount of *gamma-glutamyl transferase* |
| PROT (protein) | Amount of protein |

3. Data Preparation

The preparation of data in this study involves the process of normalizing the dataset to ensure consistency of format and standardization of values. The hepatitis C dataset used was categorized into five groups based on the patient's medical condition, namely Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis, and Cirrhosis. For the purpose of analysis, the categories were regrouped into two main classes of patients not infected with Hepatitis C (non-infected), consisting of Blood Donor (healthy blood donor) and Suspect Blood Donor (blood donation with risk indications) and patients infected with Hepatitis C (infected) which include Hepatitis (active phase of infection), Fibrosis (early-stage liver tissue damage), and Cirrhosis (advanced liver damage). This division is based on clinical considerations, where Blood Donor and Suspect Blood Donor represent subjects without a diagnosis of Hepatitis C, while the other three categories indicate disease progressivity. This approach allows for clearer analysis in distinguishing between healthy and infected patients, while minimizing bias due to variations in disease stages. In addition, the normalization of the data facilitates the application of the classification model.

4. Modeling

At the modeling stage, the Random Forest algorithm was chosen for the classification of the hepatitis C dataset due to its ability to handle unbalanced data and heterogeneous features through a robust ensemble learning mechanism. The 80:20 data sharing consisting of training and testing as well as the application of SMOTE + NearMiss resampling techniques was carried out to overcome the problem of class imbalance while validating the generalization ability of the model. Evaluations using Precision, Recall, F1-score, and AUC-ROC were comprehensively selected to measure prediction accuracy, case detection sensitivity (Recall), performance balance on unbalanced data (F1-score), and

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

modelability (AUC-ROC), all of which are crucial in the context of medical diagnoses that require high accuracy and minimal misclassification.

a. Random Forest

Random Forest is an ensemble algorithm that uses Decision Tree as the basis for which it has been developed. Ensemble algorithms are methods that combine multiple machine learning models to produce better predictions (Chowdhury dkk., 2024; Sari dkk., 2023). Random Forest performs the classification by integrating trained trees on a sample of data. The root, also known as the root, is the top node in the decision tree. Internal nodes are data that flows to the next node until it reaches the leaf node, usually branches have at least two exits and one entrance. Whereas the leaf node is the last node that has only one entrance and has no exit (Cecep Wahyu Cahyana & Akhsin Nurlayli, 2023; Kohsasih & Situmorang, 2022). The following is the formula for calculating the entropy value and the information gain value in equation 1 (Sandag, 2020).

$$Entropy\ (Y) = -\sum_i p\ (c|Y) log^2 p(c|Y), \quad (1)$$

Information:
Y = Case set
P(c|Y) = Proportion of Y value to class c

$$Information\ Gain\ (Y) = Entropy\ (Y) \sum_{vEValues(a)} \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v) \quad (2)$$

Information:
Values ($a$) = Possible values in a case set
Yv = Subclass of Y with class v which corresponds to class a
Ya = All values that correspond to a

b. SMOTE (Synthetic Minority Oversampling Technique)

Methods for handling class imbalances in a dataset. An imbalance occurs when the proportion of one class (majority) significantly exceeds the other class (minority), thus potentially causing a bias in the machine learning model against the majority class (Wang dkk., 2025). SMOTE produces synthetic samples $x_{new}$ with linear interpolation between two selected minority samples $x_i$ and $x_j$ with equation 3 as follows (Shamshuzzoha dkk., 2025).

$$x_{new} = x_i + \lambda * (x_{ij} - x_i) \quad (3)$$

Information:
$x_i$ = A randomly selected sample of the original minority from the dataset.
$\lambda$ = Random weight in intervals [0,1]
$x_{ij}$ = One of the closest k-neighbors

SMOTE effectively enhances the generalization of the model by enriching the variation of a minority sample, although it has the potential to introduce noise if the nearest neighbor is an outlier.

c. ENN (Edited Nearest Neighbors)

The undersampling method falls under the category of data cleaning methods that aims to clean up the noise in the dataset, remove ambiguous samples that are near the decision boundary, improve classification accuracy by simplifying data distribution (Mooijman dkk., 2023).

d. OSS (*One-Sided Selection*)

Undersampling technique to deal with the problem of class imbalance with a two-stage approach to clean up noise using Tomek Links and majority sample selection based on Condensed Nearest Neighbor Rule (CNN). The OSS method is particularly effective for the use of datasets with high noise levels and the majority class dominates significantly (Mooijman dkk., 2023).

5. Evaluation

Conduct a comprehensive evaluation of the model that has been built to ensure that the model is able to achieve the research objectives that have been set. The stage to test matrix confusion with the use of 80% training data and

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

test data of 20% of patients diagnosed healthy (540) was included in class 0. Patients diagnosed with infection (75) were put in class 1. In the evaluation stage, a confusion matrix is used which aims to provide a visual overview of the performance of the classification model in machine learning (Ahmed dkk., 2025). The following is the confusion matrix table in Table 2.

Table 2. Confusion Matrix

| CONFUSION MATRIX | | Class 1 | Class 2 |
|---|---|---|---|
| | | Predicted | Predicted |
| Class 1 | Actual | TP | FN |
| Class 2 | Actual | FP | TN |

The following is an explanation of the class in the confusion matrix:
a. True Positive (TP) is the data correctly predicted by the model on the positive class
b. False Positive (FP) is data that is predicted positive but not in a positive class (predicted positive but negative data)
c. True Negative (TN) is data that is incorrectly predicted by the model in the negative class
d. False Negative (FN) is data that is predicted to be negative but not in the negative class (predicted negative but positive data)

The following calculation of Accuracy, Precision and Recall in the confusion matrix is as follows:
a. Accuracy is the proportion of the correct prediction compared to the total number of predictions Equation 4

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

b. Precision is the proportion of positive class predictions out of the total positive predictions made in equation 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

c. Recall is to measure the proportion of positive samples successfully identified by the model in equation 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

d. F-Measure is Recall is a measure of the proportion of positive samples successfully identified by the model in equation 7.

$$F - Measure = 2\ \frac{Precsion * Recall}{Precision + Recall} \quad (7)$$

6. Deployment

The stage that transforms the modeling results into an operational solution that is integrated with the production system. This stage involves not only technical implementation, but also governance, monitoring, and maintenance aspects to ensure the sustainability of the solution (Schröer dkk., 2021). The deployment phase will focus on planning the implementation of the model to the application project which includes planning, monitoring, and maintenance to ensure the model remains relevant and delivers accurate results. It is hoped that this process can provide an accurate and relevant model.

## RESULT

The data analysis conducted in this study confirms that the results obtained are in line with the research objectives that have been formulated previously, while being supported by relevant theoretical foundations. The data collection process began with the collection of hepatitis C datasets from the Kaggle.com platform, which consisted of 615 data samples. Furthermore, the data preparation stage (data preprocessing) is carried out including data cleaning, transformation, and normalization to ensure optimal data quality. In the data processing process, the Random Forest (RF) method was applied with two resampling approaches, namely SMOTE + ENN (Synthetic Minority Over-

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

sampling Technique + Edited Nearest Neighbors) & SMOTE + OSS (Synthetic Minority Over-sampling Technique + One-Sided Selection) as a comparator, and experiments without resampling techniques were also carried out to evaluate the effect of data balancing on model performance. All calculations and algorithm implementations are carried out using the Python programming language by utilizing libraries such as scikit-learn, imbalanced-learn, and pandas to ensure accuracy and computational efficiency.

Table 3. Hepatitis C Dataset

| No | Category | Age | Sex | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0=Blood Donor | 32 | m | 38,5 | 52,5 | 7,7 | 22 | 7,5 | 6,93 | 3,23 | 106 | 12,1 | 69 |
| 2 | 0=Blood Donor | 32 | m | 38,5 | 70,3 | 18 | 25 | 3,9 | 11,2 | 4,8 | 74 | 15,6 | 76,5 |
| 3 | 0=Blood Donor | 32 | m | 46,9 | 74,7 | 36,2 | 53 | 6,1 | 8,84 | 5,2 | 86 | 33,2 | 79,3 |
| 4 | 0=Blood Donor | 32 | m | 43,2 | 52 | 30,6 | 23 | 18,9 | 7,33 | 4,74 | 80 | 33,8 | 75,7 |
| 5 | 0=Blood Donor | 32 | m | 39,2 | 74,1 | 32,6 | 25 | 9,6 | 9,15 | 4,32 | 76 | 29,9 | 68,7 |
| .. | …. | .. | . | … | … | … | … | … | … | … | … | … | … |
| .. | …. | .. | . | … | … | … | … | … | … | … | … | … | … |
| 613 | 3=Cirrhosis | 64 | f | 29 | 87,3 | 3,5 | 99 | 1,66 | 3,63 | 66,7 | 64,2 | 82 | 1,66 |
| 614 | 3=Cirrhosis | 46 | f | 33 | 0 | 39 | 62 | 3,56 | 4,2 | 52 | 50 | 71 | 3,56 |

**Random Forest Testing Without Resampling**

Fig 2 shows the confusion matrix produced by the matrix with a description of the model's performance in predicting the actual class based on test data without the application of resampling techniques. The Vertical (Actual) axis shows the real class of the data 0=Donor Blood (107 samples), 1=Hepatitis (5 samples), 2=Fibrosis (4 samples) and 3=Cirrhosis (6 samples). Class 0 (Blood Donor) resulted in 107 samples correctly predicted as Blood Donor (True Positive/TP), No classification errors (False Negative/FN = 0) showing accuracy (100%) for this class, showing an excellent model in identifying healthy subjects. Class 1 (Hepatitis) produced 5 actual samples of Hepatitis, but all were incorrectly classified as Blood Donor (False Negative/FN = 5), True Positive = 0 which shows that the model failed to detect cases of Hepatitis. This error indicates a majority bias due to a class imbalance, where the Blood Donor class dominates. Class 2 (Fibrosis) resulted in 4 samples correctly predicted as Fibrosis (TP = 4). There are no classification errors (FN = 0). The performance is good, but it should be noted that the small number of samples of only 4 affects the generalization of the results. Class 3 (Cirrhosis) 5 samples were correctly predicted to be Cirrhosis (TP = 5). 1 sample was misclassified as Fibrosis (FN = 1).

In Fig 3 the ROC (Receiver Operating Characteristic) curve of the Random Forest model without resampling in the Hepatitis C dataset shows a perfect classification performance separately for all classes, with an AUC (Area Under the Curve) value of 1.00 for each class, including Blood Donor (0), suspect Blood Donor (0s), Hepatitis (1), Fibrosis (2), and Cirrhosis (3). These results indicate that the model is able to perfectly separate all classes based on different classification thresholds, where the true positive rate (TPR) reaches 1.0 in the absence of false positives (FPR = 0) for each class. However, these findings are in contrast to the results of the previous confusion matrix which showed the failure of the model in classifying hepatitis cases, thus raising questions about the potential overfitting or incompatibility between ROC-based evaluations and the reality of data distribution. This disparity may be due to the dominance of the majority class (Blood Donor) which affects aggregate metrics such as AUC, while the minority class (Hepatitis) remains undetected in practice.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

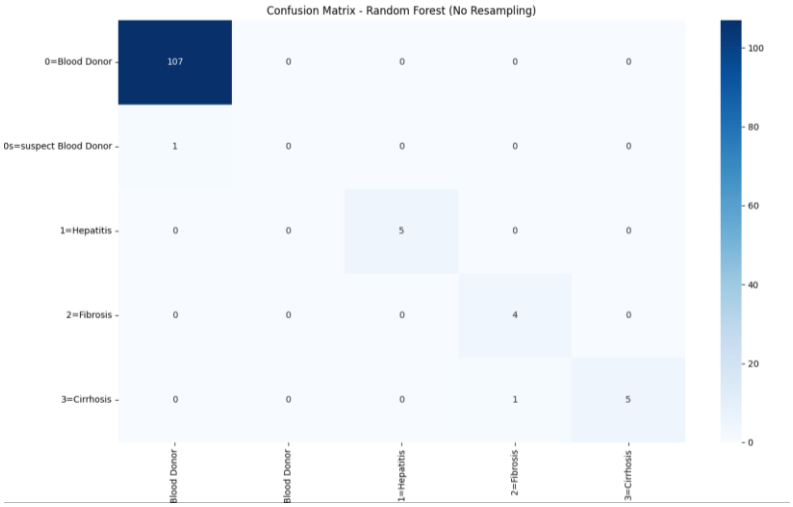



Fig. 2 Confusion Matrix Random Forest Without Resampling

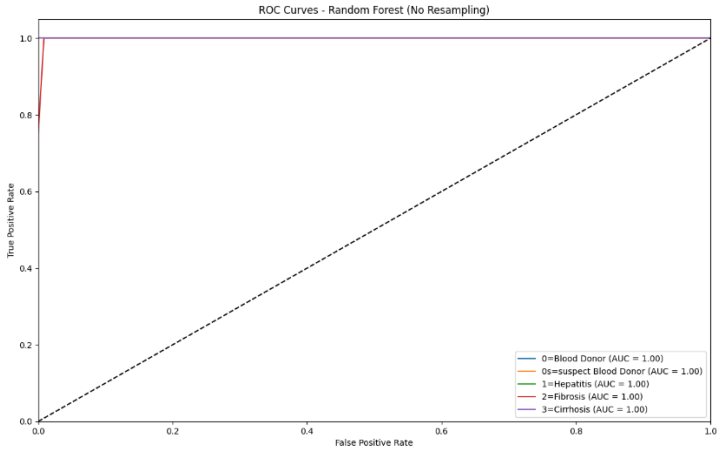



Fig. 3 Curve ROC Random Forest Without Resampling

```
=== Random Forest (No Resampling) Performance ===
Accuracy: 0.9837
Precision: 0.9773
Recall: 0.9837
F1-Score: 0.9798
ROC AUC: 0.9998

Classification Report:
                    precision    recall  f1-score   support

        0=Blood Donor       0.99      1.00      1.00       107
0s=suspect Blood Donor       0.00      0.00      0.00         1
          1=Hepatitis       1.00      1.00      1.00         5
           2=Fibrosis       0.80      1.00      0.89         4
          3=Cirrhosis       1.00      0.83      0.91         6

             accuracy                           0.98       123
            macro avg       0.76      0.77      0.76       123
         weighted avg       0.98      0.98      0.98       123
```

Fig. 4 Random Forest Evaluation Model Without Resampling

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

The results of the evaluation of the Random Forest model without the resampling technique on the Hepatitis C dataset showed high overall accuracy (0.9837) with a near-perfect ROC AUC value (0.9998), but an in-depth analysis of the classification report revealed some critical weaknesses. The model achieved perfect performance in identifying the majority class of Blood Donors (precision=0.99, recall=1.00), but failed miserably in detecting the minority class of suspect Blood Donors (precision=0.00, recall=0.00) that had only 1 sample, indicating the severe impact of the class imbalance. Meanwhile, the model shows perfect precision (1.00) for the Hepatitis class (5 samples) but with a recall of 1.00 which needs to be interpreted carefully considering the very small sample size. The Fibrosis (4 samples) and Cirrhosis (6 samples) classes achieved high recalls (1.00 and 0.83 respectively), but lower precision in Fibrosis (0.80) indicated the presence of false positives. The significant difference between macro avg (0.76) and weighted avg (0.98) confirms the model's bias towards the majority class, where the weighted aggregate metric is inflated by the dominance of the Blood Donor class.

**Random Forest Testing with SMOTE+ENN**

In Fig 5, the confusion matrix from the Random Forest model with the SMOTE+ENN resampling technique on the Hepatitis C dataset showed a significant improvement in handling class imbalances compared to the model without resampling. This model successfully classified 106 out of 107 cases of Blood Donor (0) correctly (true) positive, while only 1 case was incorrectly predicted. For the Hepatitis class (1), the model achieved 4 true positives without false negatives, indicating success in detecting previously unidentified hepatitis cases. The Fibrosis (2) and Cirrhosis (3) classes also performed well with 5 and 6 true positives respectively, although there were slight misclassifications between the two classes. These results confirm the effectiveness of SMOTE+ENN in balancing the data distribution and improving the sensitivity of the model to minority classes, while maintaining accuracy in the majority class.

In Fig 6, the ROC (Receiver Operating Characteristic) curve from the Random Forest model using the SMOTE+ENN resampling technique on the Hepatitis C dataset shows a perfect classification performance for all classes, with an AUC (Area Under the Curve) value of 1.00 for each category, including Blood Donor (0s), Suspected Blood Donor (0s), Hepatitis (1), Fibrosis (2), and Cirrhosis (3). These results indicate that the model is able to achieve a maximum true positive rate (TPR) (1.0) without producing false positives (FPR = 0) across all classification thresholds, reflecting the ideal differentiating ability between the classes.

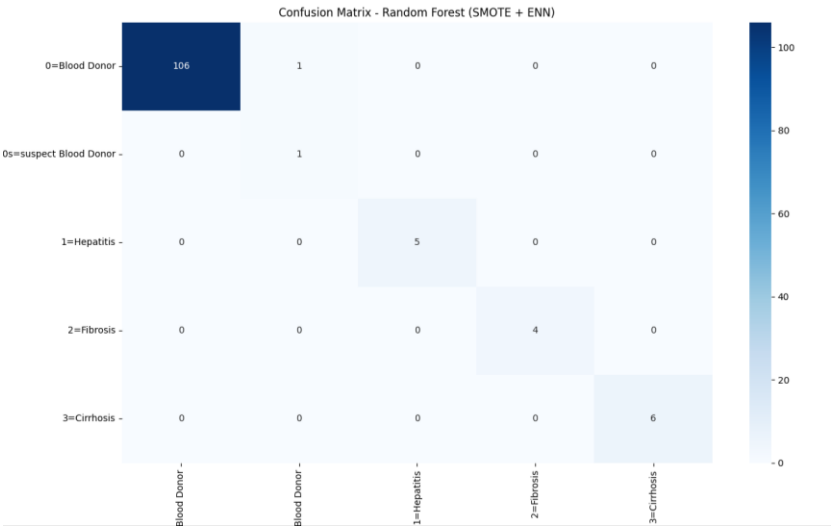


Fig. 5 Confusion Matrix Random Forest SMOTE+ENN

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

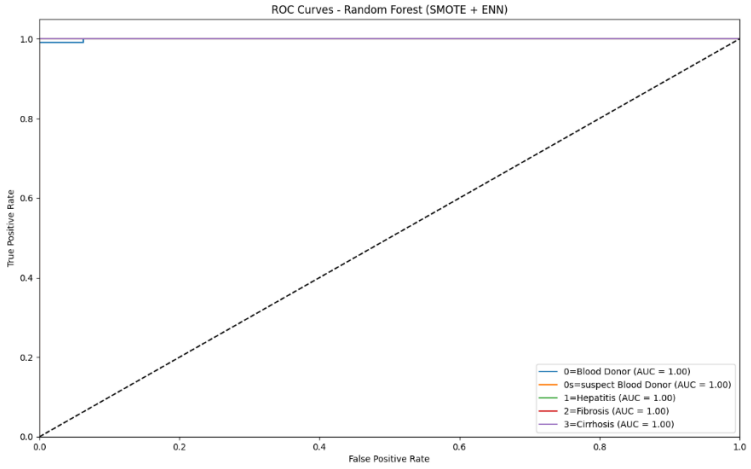



Fig. 6 Curve ROC Random Forest SMOTE+ENN

```
=== Random Forest (SMOTE + ENN) Performance ===
Accuracy: 0.9919
Precision: 0.9959
Recall: 0.9919
F1-Score: 0.9932
ROC AUC: 0.9999

Classification Report:
                    precision    recall  f1-score   support

        0=Blood Donor       1.00      0.99      1.00       107
0s=suspect Blood Donor       0.50      1.00      0.67         1
          1=Hepatitis       1.00      1.00      1.00         5
           2=Fibrosis       1.00      1.00      1.00         4
          3=Cirrhosis       1.00      1.00      1.00         6

             accuracy                           0.99       123
            macro avg       0.90      1.00      0.93       123
         weighted avg       1.00      0.99      0.99       123
```

Fig. 7 Model Evaluation Random Forest SMOTE+ENN

In Fig. 7, the results of the evaluation of the Random Forest model using the SMOTE+ENN resampling technique on the Hepatitis C dataset show a significant improvement in performance compared to the model without resampling. The model achieved very high overall accuracy (0.9919) with near-perfect AUC ROC values (0.9999), as well as improvements on all evaluation metrics, including precision (0.9959), recall (0.9919), and F1-score (0.9932). The classification report analysis revealed that the model successfully solved the class imbalance problem, with perfect performance (precision and recall = 1.00) for the Hepatitis (1), Fibrosis (2), and Cirrhosis (3) classes, as well as a precision of 1.00 and a recall of 0.99 for the majority of Blood Donor classes (0). The most noticeable increase was seen in the minority class of suspect Blood Donor (0s), which was previously undetected (recall = 0.00), now reaching a recall of 1.00, albeit with a precision of 0.50 due to limited samples (only 1 instance). The high avg macro values (0.90 for precision, 1.00 for recall, and 0.93 for F1-score indicate a better balance of performance between classes compared to non-resampling models, while confirming the effectiveness of SMOTE+ENN in handling data imbalances.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

**Random Forest Testing with SMOTE+OSS**

In Fig. 8, the confusion matrix from the Random Forest model with the SMOTE+OSS resampling technique on the Hepatitis C dataset shows excellent performance in overcoming the class imbalance problem. This model managed to classify all 107 Blood Donor samples (0) correctly (true positive), without producing false positives for this class. For the minority class, the model achieved a perfect recall (1.00) in Hepatitis (1) with 5 true positives, Fibrosis (2) with 4 true positives, and Cirrhosis (3) with 6 true positives, in the absence of false negatives. These results show that the combination of SMOTE (oversampling) and OSS (undersampling) effectively balances the distribution of the data while maintaining the original characteristics of the dataset, so that the model is able to recognize the patterns of all classes well. The absence of misclassification between classes indicates that this technique is successful in separating decision boundaries clearly, even for clinically similar classes such as Fibrosis and Cirrhosis.

In Fig. 9, the ROC (Receiver Operating Characteristic) curve from the Random Forest model using the SMOTE+OSS resampling technique on the Hepatitis C dataset shows a perfect classification performance for all classes, with an AUC (Area Under the Curve) value of 1.00 for each category, including Blood Donor (0), suspect Blood Donor (0s), Hepatitis (1), Fibrosis (2), and Cirrhosis (3). These results indicate that the model is able to achieve a maximum true positive rate (TPR) (1.0) without producing false positives (FPR = 0) at all classification thresholds, reflecting the ideal differentiating ability between classes. This consistency of perfect AUC values is in line with the results of the previous confusion matrix, where the model showed perfect precision and recall for all classes, including minority classes such as Hepatitis and suspect Blood Donor.

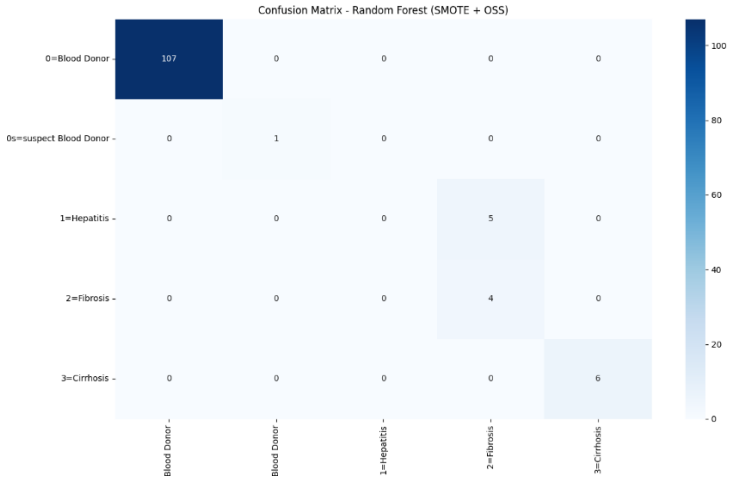



Fig. 8 Confusion Matrix Random Forest SMOTE+OSS

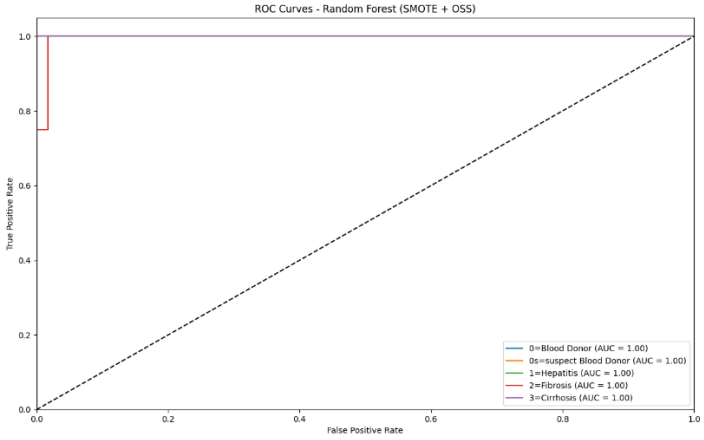



Fig. 9 Curve ROC Random Forest SMOTE+OSS

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

```
================================================
= RANDOM FOREST WITH SMOTE + ONE-SIDED SELECTION =
================================================

=== Random Forest (SMOTE + OSS) Performance ===
Accuracy: 0.9593
Precision: 0.9413
Recall: 0.9593
F1-Score: 0.9468
ROC AUC: 0.9992

Classification Report:
                    precision    recall  f1-score   support

         0=Blood Donor       1.00      1.00      1.00       107
0s=suspect Blood Donor       1.00      1.00      1.00         1
           1=Hepatitis       0.00      0.00      0.00         5
            2=Fibrosis       0.44      1.00      0.62         4
           3=Cirrhosis       1.00      1.00      1.00         6

              accuracy                           0.96       123
             macro avg       0.69      0.80      0.72       123
          weighted avg       0.94      0.96      0.95       123
```

Fig. 10 Model Evaluation Random Forest SMOTE+OSS

The results of the evaluation of the Random Forest model using the SMOTE + One-Sided Selection (OSS) resampling technique on the Hepatitis C dataset showed contradictory performance between the majority and minority classes. Although the model achieved high accuracy (0.9593) and near-perfect AUC ROC values (0.9992), classification analysis revealed a total failure in detecting cases of Hepatitis (precision=0.00, recall=0.00) as well as low precision (0.44) in the Fibrosis class despite the perfect recall (1.00). The model showed perfect performance (precision=1.00, recall=1.00) for the Blood Donor (majority), suspect Blood Donor, and Cirrhosis classes, but the significant disparity between the macro average (precision=0.69, recall=0.80) and weighted average (precision=0.94, recall=0.96) indicated a strong bias towards the majority class. These results highlight the limitations of SMOTE+OSS in handling extreme minority class cases (such as Hepatitis with only 5 samples), where OSS undersampling techniques may be too aggressive in removing important instances.

## DISCUSSIONS

The comprehensive evaluation of Random Forest models on the Hepatitis C dataset (obtained from Kaggle) reveals significant insights into the impact of different resampling strategies on classification performance for this clinically important multi-class problem. Our analysis compares three distinct approaches baseline Random Forest without resampling, Random Forest with SMOTE+ENN, and Random Forest with SMOTE+OSS, each demonstrating unique characteristics in handling the inherent class imbalance. The baseline Random Forest model without resampling achieved strong overall performance (Accuracy=0.9837, ROC AUC=0.9998), yet exhibited critical limitations in minority class detection. While demonstrating perfect recall for the majority Blood Donor class (1.00) and Hepatitis cases (1.00), it completely failed to identify suspect Blood Donor instances (recall=0.00) and showed suboptimal precision for Fibrosis cases (0.80). The substantial disparity between macro average (precision=0.76) and weighted average (precision=0.98) metrics clearly illustrates the model's bias toward the majority class, a well-documented challenge in medical diagnostic applications where minority classes often carry critical clinical significance.

Implementation of SMOTE+ENN resampling yielded marked improvements across all evaluation metrics (Accuracy=0.9919, ROC AUC=0.9999). This approach successfully addressed the model's previous failure with suspect Blood Donor cases (recall improved from 0.00 to 1.00) while maintaining excellent performance for other classes. Particularly noteworthy is the perfect classification achieved for Hepatitis, Fibrosis, and Cirrhosis cases (precision=1.00, recall=1.00), suggesting that the combination of synthetic minority oversampling and edited nearest neighbor cleaning effectively balanced the dataset without introducing significant noise. The improved macro average metrics (precision=0.90, recall=1.00) demonstrate this technique's superiority in creating a more robust and generalizable model for clinical applications.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

The SMOTE+OSS approach presented more nuanced results (Accuracy=0.9593, ROC AUC=0.9992). While maintaining perfect classification for Blood Donor, suspect Blood Donor, and Cirrhosis classes, it completely failed to detect Hepatitis cases (recall=0.00) and showed poor precision for Fibrosis (0.44) despite perfect recall (1.00). These results suggest that the one-sided selection undersampling may have been overly aggressive, potentially removing critical instances needed for accurate minority class identification. The approach's limitations are particularly evident in the macro average precision (0.69), which remains substantially lower than the weighted average (0.94), indicating persistent challenges with class imbalance. The following table 4 is a comparison of the results of the classification report

Table 4. Comparison of Classification Report Results

| Metric | Random Forest (No Resampling) | Random Forest (SMOTE + ENN) | Random Forest (SMOTE + OSS) |
|---|---|---|---|
| Accuracy | 0.9837 | 0.9919 | 0.9593 |
| Precision | 0.9773 | 0.9959 | 0.9413 |
| Recall | 0.9837 | 0.9919 | 0.9593 |
| F1-Score | 0.9798 | 0.9932 | 0.9468 |
| ROC AUC | 0.9998 | 0.9999 | 0.9992 |

## CONCLUSION

This study demonstrates that resampling strategies significantly impact the performance of Random Forest models in classifying Hepatitis C data, with SMOTE+ENN emerging as the most effective approach. While the baseline model without resampling showed bias toward majority classes (Blood Donor accuracy=1.00, but suspect Blood Donor recall=0.00), SMOTE+ENN successfully addressed class imbalance by achieving perfect classification (precision=1.00, recall=1.00) for key minority classes (Hepatitis, Fibrosis, Cirrhosis) and improving suspect Blood Donor detection (recall=1.00). In contrast, SMOTE+OSS showed limitations in identifying Hepatitis cases (recall=0.00) and Fibrosis precision (0.44), suggesting its undersampling approach may be too aggressive for medical diagnostic applications. These findings highlight that proper resampling technique selection is crucial for developing reliable diagnostic tools, with SMOTE+ENN proving particularly effective for handling class imbalance in Hepatitis C classification while maintaining high overall performance (Accuracy=0.9919, ROC AUC=0.9999). Future research should focus on optimizing resampling techniques for extreme class imbalance scenarios common in medical datasets.

## REFERENCES

Ahmed, K. H., Axelsson, S., Li, Y., & Sagheer, A. M. (2025). A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling. *Machine Learning with Applications*, *20*, 100675. https://doi.org/10.1016/j.mlwa.2025.100675

Baskara, A. A., Piranti, N. M., & Romdendine, M. F. (2025). FRAMEWORK DATA MINING: SEBUAH SURVEI. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(3), Article 3. https://doi.org/10.36040/jati.v9i3.13803

Bennin, K. E., Tahir, A., MacDonell, S. G., & Börstler, J. (2022). An empirical study on the effectiveness of data resampling approaches for cross-project software defect prediction. *IET Software*, *16*(2), 185–199. https://doi.org/10.1049/sfw2.12052

Canero, F. M., Rodriguez-Galiano, V., & Aragones, D. (2024). Machine Learning and Feature Selection for soil spectroscopy. An evaluation of Random Forest wrappers to predict soil organic matter, clay, and carbonates. *Heliyon*, *10*(9), e30228. https://doi.org/10.1016/j.heliyon.2024.e30228

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

Cecep Wahyu Cahyana & Akhsin Nurlayli. (2023). Analisis Performa Logistic Regression, Naïve Bayes, dan Random Forest sebagai Algoritma Pendeteksi Kanker Payudara. *INSERT : Information System and Emerging Technology Journal*, *4*(1), 51–64. https://doi.org/10.23887/insert.v4i1.62362

Charisma, R. A., Pamungkas, S., Saputra, R. A., Ramadhan, N. G., & Adhinata, F. D. (2022). Analisis Penerapan Metode Ensembled Learning Decision Tree Pada Klasifikasi Virus Hepatitis C. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 405–409. https://doi.org/10.47065/josyc.v3i4.2064

Chen, K., Abtahi, F., Carrero, J.-J., Fernandez-Llatas, C., & Seoane, F. (2023). Process mining and data mining applications in the domain of chronic diseases: A systematic review. *Artificial Intelligence in Medicine*, *144*, 102645. https://doi.org/10.1016/j.artmed.2023.102645

Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*, *5*, 100297. https://doi.org/10.1016/j.health.2023.100297

Diniarti, F., & Romli, H. (2024). PENULARAN INFEKSI HEPATITIS B MELALUI PASANGAN SEKSUAL PADA IBU HAMIL DI KOTA BENGKULU, INDONESIA. *Journal of Nursing and Public Health*, *Vol 12 No 2 (2024)*. https://jurnal.unived.ac.id/index.php/jnph/article/view/7444

Elliott, S., Flynn, E., Mathew, S., Hajarizadeh, B., Martinello, M., Wand, H., & Ward, J. (2024). Prevalence of hepatitis C virus exposure and infection among Indigenous and tribal populations: A global systematic review and meta-analysis. *Public Health*, *233*, 65–73. https://doi.org/10.1016/j.puhe.2024.04.035

Hakim, D. K., Gernowo, R., & Nirwansyah, A. W. (2024). Flood prediction with time series data mining: Systematic review. *Natural Hazards Research*, *4*(2), 194–220. https://doi.org/10.1016/j.nhres.2023.10.001

Iman, Q., & Wijayanto, A. W. (2021). Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera (Rastra) di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine. *Jurnal Sistem dan Teknologi Informasi (Justin)*, *9*(2), 178. https://doi.org/10.26418/justin.v9i2.44137

Kohsasih, K. L., & Situmorang, Z. (2022). Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular. *Jurnal Informatika*, *9*(1), 13–17. https://doi.org/10.31294/inf.v9i1.11931

Kurnianingsih, Nobukawa, S., Widyawati, M. N., Pramana, C., Aji, N. B., Thohari, A. N. A., Hendrawati, D., Sato-Shimokawara, E., & Kubota, N. (2025). A Novel Ensemble XGBoost and Deep Q-Network for Pregnancy Risk Prediction on Multi-Class Imbalanced Datasets. *ICT Express*, S2405959525000724. https://doi.org/10.1016/j.icte.2025.05.010

Liu, P., Liu, Y., Liu, H., Xiong, L., Mei, C., & Yuan, L. (2024). A Random Forest Algorithm for Assessing Risk Factors Associated With Chronic Kidney Disease: Observational Study. *Asian/Pacific Island Nursing Journal*, *8*, e48378. https://doi.org/10.2196/48378

Mehraeen, E., Janfaza, N., Shahidi, R., Afzalian, A., Varshochi, S., Jashaninejad, R., Pashaei, A., Tantuoyir, M. M., Camara, M., Paranjkhoo, P., Parmoon, Z., Elnaz, S., Salmani, R., Matini, P., Mirzapour, P., Ebrahimi, H., Moradi, A., SeyedAlinaghi, S., & Jahanfar, S. (2024). Important Risk Factors of Liver Cirrhosis in HIV and Hepatitis C Coinfected Patients: A Systematic Review. *The Open AIDS Journal*, *18*(1), e18746136280350. https://doi.org/10.2174/0118746136280350240214064332

Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A., & Blokland, M. (2023). The effects of data balancing approaches: A case study. *Applied Soft Computing*, *132*, 109853. https://doi.org/10.1016/j.asoc.2022.109853

Nie, S., Chen, H., Sun, X., & An, Y. (2024). Spatial Distribution Prediction of Soil Heavy Metals Based on Random Forest Model. *Sustainability*, *16*(11), 4358. https://doi.org/10.3390/su16114358

Pourkarim, M. R. (2022). World Hepatitis Day in 2022: Challenges of Viral Hepatitis Elimination in Elongated COVID-19 Pandemic. *Pathogens*, *11*(9), 1002. https://doi.org/10.3390/pathogens11091002

Sandag, G. A. (2020). Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. *CogITo Smart Journal*, *6*(2), 167–178. https://doi.org/10.31154/cogito.v6i2.270.167-178

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, *14*(1), 155–162. https://doi.org/10.35970/infotekmesin.v14i1.1751

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Schwarz, T., Horváth, I., Fenz, L., Schmutterer, I., Rosian-Schikuta, I., & Mårdh, O. (2022). Interventions to increase linkage to care and adherence to treatment for hepatitis C among people who inject drugs: A systematic

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6089

**Submitted** : Jun 13, 2025
**Accepted** : Jul 02, 2025
**Published** : Jul 05, 2025

review and practical considerations from an expert panel consultation. *International Journal of Drug Policy*, *102*, 103588. https://doi.org/10.1016/j.drugpo.2022.103588

Shamshuzzoha, Md., Audry, T. T. B., Alam, Md. J., Bhuiyan, Z. A., Motaharul Islam, M., & Hassan, M. M. (2025). A novel framework for seasonal affective disorder detection: Comprehensive machine learning analysis using multimodal social media data and SMOTE. *Acta Psychologica*, *256*, 105005. https://doi.org/10.1016/j.actpsy.2025.105005

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. *SBC Reviews on Computer Science*, *4*(1), 28–43. https://doi.org/10.5753/reviews.2024.3757

Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, *41*, 101316. https://doi.org/10.1016/j.imu.2023.101316

Sutradhar, A., Akter, S., Shamrat, F. M. J. M., Ghosh, P., Zhou, X., Idris, M. Y. I. B., Ahmed, K., & Moni, M. A. (2024). Advancing thyroid care: An accurate trustworthy diagnostics system with interpretable AI and hybrid machine learning techniques. *Heliyon*, *10*(17), e36556. https://doi.org/10.1016/j.heliyon.2024.e36556

Wang, A. X., Le, V.-T., Trung, H. N., & Nguyen, B. P. (2025). Addressing imbalance in health data: Synthetic minority oversampling using deep learning. *Computers in Biology and Medicine*, *188*, 109830. https://doi.org/10.1016/j.compbiomed.2025.109830

Zhang, Y., Tapa, J., Johnson, C. C., Phillips, T. R., Fairley, C. K., Ameyan, W., Mello, M. B., Chow, E. P. F., Chidarikire, T., & Ong, J. J. (2025). HIV, hepatitis, and syphilis self-testing among adolescents and young adults: A systematic review and meta-analysis. *Journal of Infection and Public Health*, *18*(6), 102764. https://doi.org/10.1016/j.jiph.2025.102764

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.