



Comparative Analysis of Random Forest, Explainable Boosting Machine and Ensemble Stacking Performance for Hepatitis C Disease Classification

Anastasia Ngeni Bagur^{1*}, Irfan Pratama²

^{1,2}Department of Information System, Mercu Buana University of Yogyakarta, Indonesia

E-Mail: ¹211210040@student.mercubuana-yogya.ac.id , ²irfanp@mercubuana-yogya.ac.id

Received Jan 06th 2026; Revised Feb 21th 2026; Accepted Mar 01st 2026; Available Online Apr 19th 2026

Corresponding Author: Anastasia Ngeni Bagur

Copyright ©2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

This study analyzed and compared the performance of three machine learning methods: Random Forest, Explainable Boosting Machine, and a Stacking Ensemble method for Hepatitis C disease classification. The study evaluated the effects of handling extreme values using the interquartile range method and applying class-balancing oversampling to the training data. A dataset of 615 patient samples, categorized into five severity classes, was used. Experiments were conducted across four scenarios: with and without outlier correction, and with and without class balancing. Model performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve. Results showed that class balancing consistently improved all macro-averaged performance metrics. The combination of Random Forest with oversampling prior to outlier correction achieved the highest F1-score of 0.8086 and an area under the curve of 0.9710. These findings highlighted the importance of addressing class imbalance to improve the recognition of minority classes in clinical datasets and demonstrated the potential of ensemble methods for reliable severity classification in Hepatitis C.

Keywords: Explainable Boosting Machine, Hepatitis C, Random Forest, SMOTE, Stacking Ensemble

1. INTRODUCTION

Hepatitis C remains a significant global public health concern as a chronic liver disease. Infection by the Hepatitis C virus (HCV) typically progresses slowly and is often asymptomatic in early stages, which results in many patients being diagnosed only after the disease has advanced to fibrosis, cirrhosis, or hepatocellular carcinoma [1]. Delayed diagnosis increases the risk of complications and mortality and imposes substantial long-term healthcare costs [2].

The integration of information technology into healthcare has promoted the use of clinical and laboratory data as the basis for computational decision support. Machine learning methods are widely adopted because they can learn complex patterns and nonlinear relationships from medical data, often yielding higher classification accuracy than conventional statistical approaches [3]. Machine learning applications in healthcare have been shown to support clinical decision-making in an objective and consistent manner [4]. In Hepatitis C cases, laboratory parameters such as liver enzyme levels, bilirubin, and hematological indicators have been reported to contribute significantly to assessing disease severity [5].

A principal challenge in applying machine learning to Hepatitis C classification is class imbalance. Patient data distributions are frequently dominated by healthy donors or early-stage cases, while advanced disease categories such as fibrosis and cirrhosis contain comparatively few samples [6]. This imbalance can cause models to favor majority classes, producing high overall accuracy but poor performance on clinically critical minority classes [7]. Consequently, macro-averaged evaluation metrics (e.g., recall, F1-score, and AUC) are more appropriate for a comprehensive assessment of model performance [8].

Medical datasets also commonly contain extreme values or outliers, which may arise from data entry errors, instrument faults, or atypical biological variation [9]. Unaddressed outliers can distort the data distribution and reduce the stability and reliability of machine learning predictions [10]. Therefore, the Interquartile Range (IQR) method is frequently employed in medical research to detect and mitigate the influence of extreme values without discarding large amounts of data [11].

Various machine learning algorithms have been applied to liver disease classification, including Random Forest, which is noted for good generalization and robustness to noise [12]. Explainable Boosting Machine (EBM) has been developed to balance predictive performance with model interpretability, a crucial consideration in clinical contexts [13]. Advanced ensemble approaches such as Stacking have been used to combine the strengths of multiple base learners to improve classification accuracy [14]. However, studies that explicitly compare Random Forest, EBM, and Stacking Ensemble while accounting for outlier treatment and class imbalance in Hepatitis C severity classification are still limited [15]. Accordingly, this research aims to analyze and compare the performance of these three methods, applying IQR for outlier handling and SMOTE to address class imbalance.

2. MATERIALS AND METHOD

This study employed a quantitative approach with an experimental machine learning framework to classify the severity levels of Hepatitis C. All research stages, from data processing to model evaluation, were implemented using Python programming language with widely used libraries for data analysis and machine learning in medical research [16]. The research workflow included data preparation, preprocessing, model training, and performance evaluation, as illustrated in Figure 1.

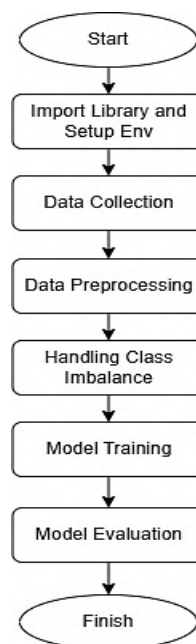


Figure 1. Research Methodology

2.1. Environment Setup and Library Import

The initial stage of the study involved preparing the working environment and importing the required libraries. The pandas and numpy libraries were utilized for data manipulation and analysis, while scikit-learn was employed for data splitting, training Random Forest and Logistic Regression models, and calculating evaluation metrics. The Explainable Boosting Machine was implemented using the interpret library, and class imbalance handling was performed using the imbalanced-learn library through the Synthetic Minority Over-sampling Technique [17].

2.2. Dataset

The dataset used in this study was obtained from Kaggle and contains clinical and laboratory data of patients. It consists of 615 samples with 12 features, namely: Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT, representing important laboratory parameters, along with one multiclass target attribute indicating patient condition. The target classes are Blood Donor (Normal), Suspect Blood Donor, Hepatitis, Fibrosis, and Cirrhosis. The class distribution is imbalanced, with the healthy donor class dominating, while advanced disease classes such as Fibrosis and Cirrhosis contain fewer samples. This imbalance necessitates special preprocessing techniques, such as oversampling, to improve the classification model's performance [18].

2.3. Data Preprocessing

Data preprocessing aimed to improve data quality and prepare it for model training. This stage involved handling missing values, converting categorical features to numerical representations using label

encoding, and standardizing numerical features with StandardScaler to ensure all features were on a comparable scale. Additionally, outliers were identified using the Interquartile Range (IQR) method, which uses quartile limits to define extreme values. This approach reduced the influence of outliers on model training without significantly removing critical information from the dataset [19]. As the final step in the preprocessing stage, the dataset was split into training and test sets at a 70:30 ratio. The training set was used to learn patterns and relationships within the Hepatitis C data, while the testing set was reserved for evaluating model performance on unseen data. The splitting process was performed randomly to maintain a representative class distribution and to reduce potential bias, thereby supporting an objective assessment of the model's generalization capability and minimizing overfitting.

2.4. Handling Class Imbalance

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique. This method generated synthetic samples for minority classes based on feature-space proximity. In this study, oversampling was applied exclusively to the training data after data splitting to prevent data leakage. This approach effectively improved the model's sensitivity to minority classes in medical datasets [20].

2.5. Classification Model Training

2.5.1. Random Forest

Random Forest is an ensemble tree-based method that constructs multiple decision trees using random subsets of data and features. Final predictions are obtained through majority voting across all trees. This model was selected for its strong generalization ability, robustness to overfitting, and capacity to capture nonlinear relationships commonly found in medical data [21].

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\} \quad (1)$$

Description:

- $\hat{y}$: The final predicted class (Hepatitis C severity level).
- Mode : The function representing the majority voting mechanism.
- $h_1(\mathbf{x})$: The prediction generated by the individual decision tree for input vector ($\mathbf{x}$)
- T : The total number of decision trees in the ensemble.
- ($\mathbf{x}$) : The input feature vector containing patient laboratory parameters.

2.5.2. Explainable Boosting Machine (EBM)

The Explainable Boosting Machine is an interpretable model belonging to the Generalized Additive Model family. It combines boosting techniques with an additive structure to deliver competitive predictive performance while providing feature-level interpretability. This interpretability is crucial in the context of Hepatitis C severity classification, supporting clinical decision-making [22].

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j) \quad (2)$$

Description:

- g : The link function (typically the logit function for classification tasks).
- $E[y]$: The expected value or probability of the target variable.
- β_0 : The intercept or baseline constant.
- $f_i(x_i)$: The shape function for a single feature x_i representing the feature's individual contribution to model.
- $f_{ij}(x_i, x_j)$: The interaction function that captures the joint effect of two features on the prediction.

2.5.3. Stacking Ensemble

The Stacking Ensemble approach combines the strengths of multiple base learners. In this study, Random Forest and EBM served as base learners, while Logistic Regression acted as the meta-learner. The probability predictions from each base learner were used as inputs for the meta-learner to generate the final prediction. This approach was expected to improve model stability and performance on complex and imbalanced datasets [23].

$$\widehat{y}_{final} = L(P_1(\mathbf{x}), P_2(\mathbf{x}), \dots, P_M(\mathbf{x})) \quad (3)$$

Description:

- $\widehat{y}_{final}$: The final class prediction produced by the ensemble system.
- L : The meta-learner function (Logistic Regression) optimizes the combination of base learner outputs.

$P_M(\mathbf{x})$: The predicted probability or output generated by the base learner (Random Forest or EBM).
 M : The total number of base learners utilized in the stacking architecture.

2.6. Model Evaluation

Model performance was evaluated using accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and the area under the receiver operating characteristic curve (AUC) using a macro averaging approach. Macro-averaged metrics were employed to provide fair evaluation across all classes, particularly for minority classes that are clinically significant [24].

1. Accuracy

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

2. Precision (Macro)

$$\text{Precision}_{\text{macro}} = \frac{1}{K} \sum_{i=1}^K \frac{TP_i}{TP_i+FP_i} \quad (5)$$

3. Recall (Macro)

$$\text{Recall}_{\text{macro}} = \frac{1}{K} \sum_{i=1}^K \frac{TP_i}{TP_i+FN_i} \quad (6)$$

4. F1-Score (Macro)

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{i=1}^K \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (7)$$

5. AUC (Macro)

$$\text{AUC}_{\text{macro}} = \frac{1}{K} \sum_{i=1}^K \text{AUC}_i \quad (8)$$

3. RESULTS AND DISCUSSION

3.1. Outlier Visualization and Data Distribution

Outlier analysis was conducted using the Interquartile Range (IQR) method. The boxplots before and after outlier correction (Figure 2) show extreme-value concentrations in several clinical features, affecting the range of variance. Out of a total of 615 samples, 168 rows were identified as containing at least one extreme value. However, in the applied handling procedure, no rows were removed, and all extreme values were corrected so that the total number of samples remained 615. This adjustment approach preserves class representation and enables a fair comparison between scenarios (before vs. after outlier handling). The distributional changes observed in the post-handling boxplot indicate reduced skewness and diminished influence of extreme values in several features, which is expected to contribute to improved stability during the model training process.

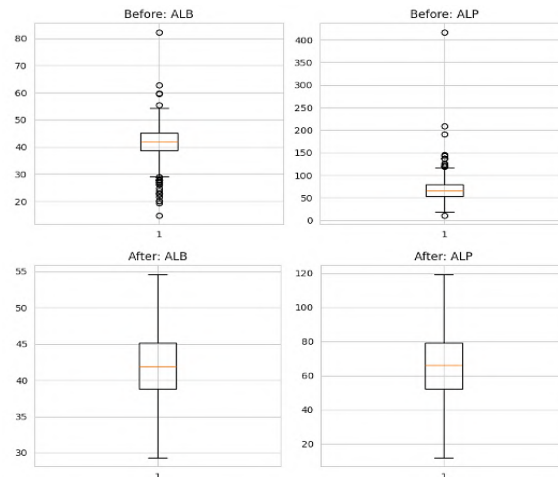


Figure 2. Boxplot of Data Before and After Outlier Handling

As a complement to the boxplot, an additional visualization outlier (Figure 3) was included. This figure presents the indices of rows containing outliers along with the number of outliers per row and per feature, facilitating identification of whether extreme values are concentrated in specific samples or distributed more evenly across the dataset. The visualization shows that a small subset of samples contains multiple extreme values, while many other samples contain only one or two values outside the IQR bounds. This pattern supports the decision to correct extreme values rather than remove entire rows, as row deletion could reduce minority class representation.

3.2. SMOTE Visualization and Class Distribution After Oversampling

The Synthetic Minority Over-sampling Technique was applied exclusively to the training data to prevent data leakage. Class distribution visualization before and after oversampling (Figure 4) showed significant transformation. The training set increased from 430 to 1,860 samples, with 1,430 synthetic samples generated by SMOTE. The increased representation of minority classes provided the model with improved opportunities to learn their characteristic patterns. The post-SMOTE bar chart demonstrated more balanced class representation, conceptually supporting improvements in macro-averaged metrics such as precision, recall, and F1-score.

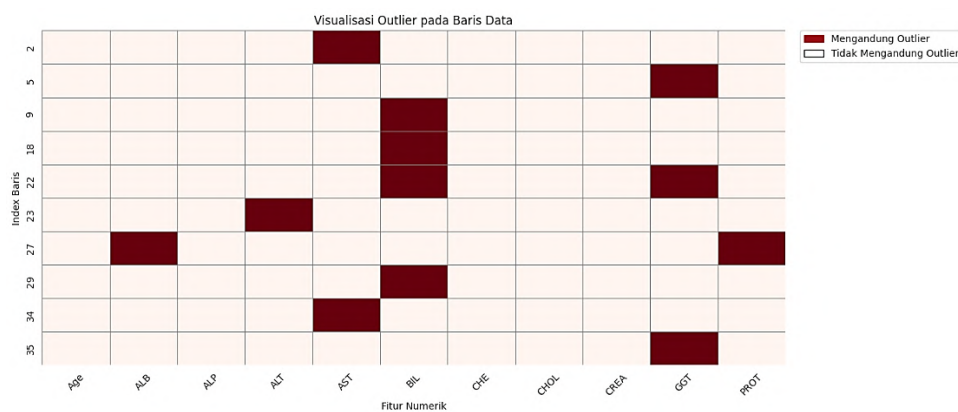


Figure 3. Visualization Outliers in Data Rows

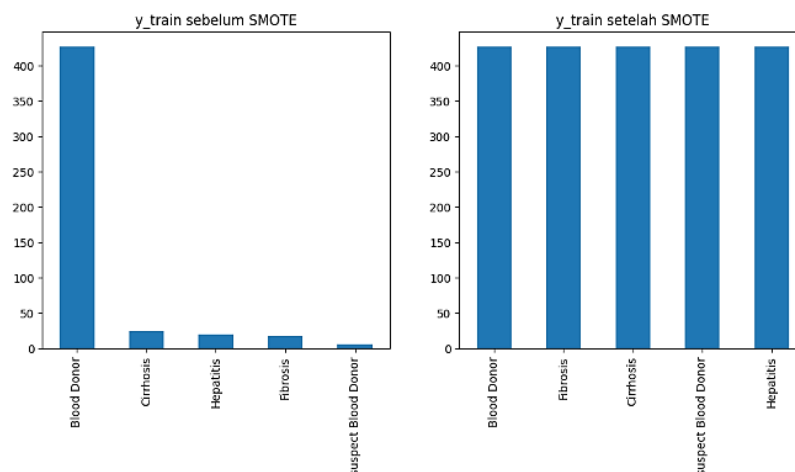


Figure 4. Class Distribution Before and After SMOTE

3.3. Results of Model Evaluation

Table 1. Before Outlier Handling

SMOTE	MODEL	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	AUC (Macro)
No	EBM	0.9351	0.5673	0.5079	0.5255	0.9729
No	RandomForest	0.9297	0.6185	0.4746	0.4866	0.9863
No	Stacking	0.9351	0.6507	0.6460	0.6246	0.9345
Yes	EBM	0.9459	0.7575	0.7610	0.7582	0.9422
Yes	RandomForest	0.9568	0.8274	0.7956	0.8086	0.9710
Yes	Stacking	0.9405	0.5873	0.5302	0.5466	0.9848

Table 2. After Outlier Handling

SMOTE	MODEL	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	AUC (Macro)
No	EBM	0.9297	0.5595	0.4857	0.5109	0.9770
No	RandomForest	0.9405	0.6061	0.5302	0.5537	0.9856
No	Stacking	0.9297	0.5584	0.6384	0.5717	0.9344
Yes	EBM	0.9459	0.7648	0.6413	0.6812	0.9456
Yes	RandomForest	0.9405	0.7106	0.6721	0.6896	0.9680
Yes	Stacking	0.9459	0.7763	0.6413	0.6879	0.9890

Tables 1 and 2 present the evaluation results of three classification models: Explainable Boosting Machine (EBM), Random Forest, and Stacking, under two data conditions (before and after outlier handling), with and without the application of SMOTE. Bold values indicate the best performance for each metric and serve as the primary basis for model performance analysis.

Before outlier handling, without SMOTE, Random Forest achieved the highest Macro AUC of 0.9863, indicating excellent class discrimination. However, the relatively low Macro Recall and F1-Score suggest that the model was unable to classify all classes evenly due to data imbalance. After applying SMOTE under the same conditions, Random Forest demonstrated the most significant improvement, achieving an accuracy of 0.9568, Macro Precision of 0.8274, Macro Recall of 0.7956, and the highest Macro F1-Score of 0.8086. These results confirm that SMOTE effectively improved class-balanced predictions, particularly for minority classes, making this configuration the best-performing overall.

Under the condition of after outlier handling with SMOTE, EBM, and Random Forest maintained relatively stable performance, although the Macro F1-Score slightly decreased compared to the scenario before outlier handling. Meanwhile, the Stacking model achieved the highest Macro AUC of 0.9890, indicating very strong class separation; however, this was not accompanied by improved class-balanced classification.

Overall, the results indicate that the application of SMOTE had a more significant impact on improving classification performance compared to outlier handling. Random Forest combined with SMOTE before outlier handling represented the most optimal configuration, producing the most balanced performance, particularly in terms of Macro F1-Score.

3.4. Effect of Outlier Handling on Model Performance

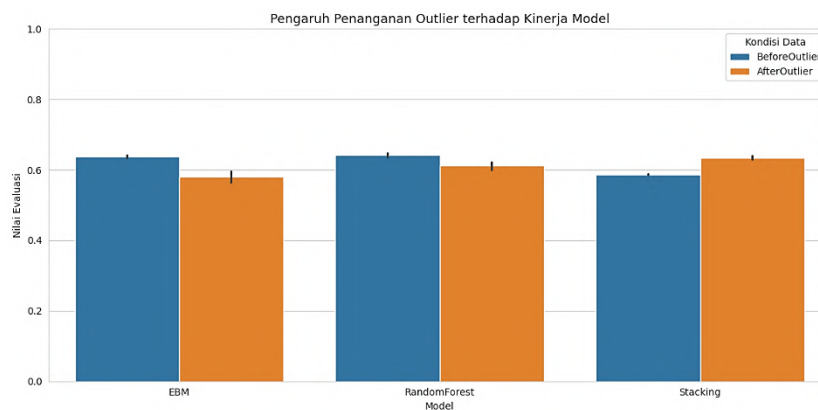


Figure 5. Outlier Handling Effect

Figure 5 shows the comparison of the average performance of the Explainable Boosting Machine (EBM), Random Forest, and Stacking models before and after outlier handling. The reported values represent macro-averaged evaluation metrics, thereby reflecting the models' balanced classification performance across all classes.

Based on the visualization, the impact of outlier handling varies across models. For EBM and Random Forest, performance after outlier treatment appears slightly lower compared to the condition without outlier correction. This finding suggests that extreme value correction does not necessarily enhance classification performance and, in certain cases, may remove informative signals that contribute to effective class discrimination.

In contrast, the Stacking model demonstrates a tendency toward improved performance following outlier handling. This indicates that the ensemble-based Stacking approach may exhibit greater adaptability to changes in data distribution, enabling it to leverage normalized data more effectively. Overall, the

visualization suggests that outlier treatment does not yield uniformly positive effects across all models; rather, its impact is contingent upon the underlying characteristics and learning mechanisms of each algorithm.

Overall, outlier handling strategies should be applied selectively and with consideration of model-specific behavior. A study conducted by Khan et al. (2024) in Expert Systems with Applications reported that, within classification contexts, outliers are not invariably detrimental [25]. In some scenarios, extreme observations may contain discriminative information essential for class separation. The study further indicated that indiscriminate removal or global correction of outliers can degrade performance in certain models due to the loss of relevant structural information. These findings are consistent with the present study, where EBM and Random Forest experienced a slight decrease in performance after outlier treatment, while the ensemble-based Stacking model demonstrated greater robustness to distributional changes.

3.5 Effect of SMOTE on Model Performance

In Figure 6, the application of SMOTE had a more pronounced impact on model performance compared to outlier handling. Under conditions without SMOTE, all three models tended to exhibit lower Macro Recall and Macro F1-Score, despite relatively high Accuracy and AUC values. This indicates that the models were biased toward predicting majority classes. After applying SMOTE, all models demonstrated improvements in Macro Precision, Macro Recall, and Macro F1-Score, both before and after outlier handling. This improvement indicates that SMOTE effectively helps the models recognize minority classes more evenly. Random Forest exhibited the most stable performance, achieving relatively high Macro F1-Score values, while Stacking excelled in Macro AUC under certain scenarios. Overall, these results suggest that data balancing through SMOTE contributes more significantly to improving classification performance than outlier handling.

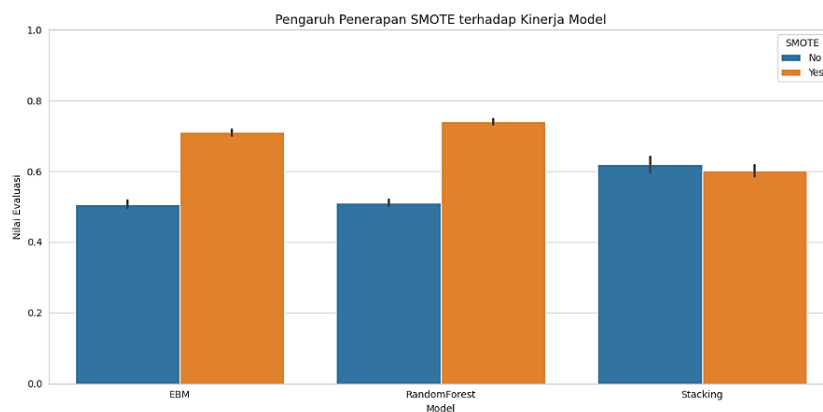


Figure 6. SMOTE Effect

3.6 Comparison of Model Performance

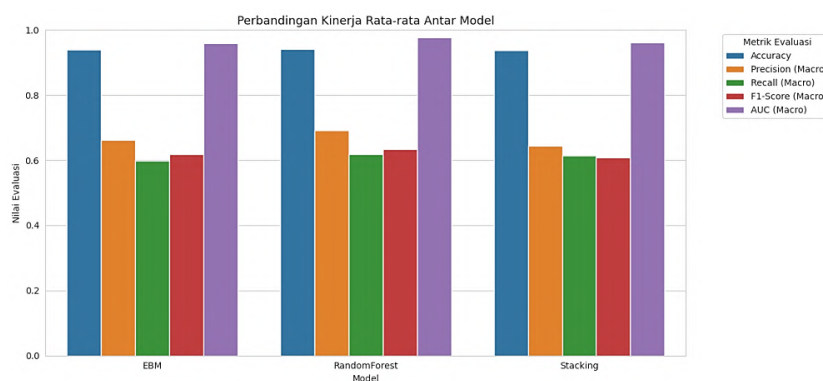


Figure 7. Model Performance Comparison

Based on the evaluation results and average performance comparison among models in Figure 7, Explainable Boosting Machine (EBM), Random Forest, and Stacking Ensemble exhibited relatively high and stable Accuracy and Macro AUC values. This indicates that all three models possess strong class discrimination capabilities in distinguishing positive and negative classes overall. However, performance

differences among the models become more apparent when analyzed using macro-averaged metrics, namely Precision, Recall, and F1-Score, which better represent the balance of performance across all classes.

Random Forest demonstrated the most consistent performance, achieving higher Macro Precision, Recall, and F1-Score than EBM and Stacking in most testing scenarios, particularly after SMOTE application. This suggests that Random Forest is more effective in producing balanced classification across all classes. EBM showed moderately stable performance with intermediate macro-metric values, while the Stacking model tended to excel in Macro AUC but had a relatively lower Macro F1-Score. This indicates that although Stacking exhibits strong class separation, its class-balanced classification is not yet optimal. Overall, Random Forest is the most balanced and stable model in this study.

3.7 Discussion of Results

The results indicate that class imbalance is a primary factor influencing the performance of classification models on the Hepatitis C dataset. This is evident in the baseline condition without SMOTE, where Accuracy and AUC values are relatively high, while macro-based metrics such as Recall and F1-score are comparatively lower. This pattern suggests that the models tend to be biased toward the majority class and are less effective in recognizing minority classes, despite demonstrating generally strong discriminative ability.

The implementation of SMOTE produced the most substantial improvement in classification performance. After applying oversampling to the training data, all models exhibited consistent increases in Macro Precision, Recall, and F1-score. These results confirm that a more balanced class distribution enables models to learn minority class patterns more effectively. This finding aligns with the study by Salmi et al. (2024), which reported that applying SMOTE to imbalanced medical datasets significantly improved recall- and F1-based metrics due to enhanced minority class representation [26]. Compared with outlier handling, SMOTE contributed more substantially to improving classification balance, particularly for the Random Forest model, which demonstrated the most stable and balanced performance.

In contrast, outlier treatment through extreme value correction did not consistently improve performance. In several scenarios, model performance slightly declined after outlier handling, indicating that certain extreme values in clinical data may still contain meaningful information. This observation is supported by Yang et al. (2024), who demonstrated that outliers in classification tasks can carry informative signals related to class structure; thus, indiscriminate removal or correction may reduce a model's ability to distinguish between classes [27]. In this study, the Stacking model exhibited greater robustness to distributional changes, as reflected by its consistently high AUC values.

Overall, these findings emphasize that preprocessing strategies should be tailored to the characteristics of both the dataset and the applied models. The combination of data balancing techniques and appropriate model selection yields a more significant improvement in classification performance than outlier handling applied in isolation.

4. CONCLUSION

This study compared the performance of Random Forest, Explainable Boosting Machine (EBM), and Stacking Ensemble in classifying the severity levels of Hepatitis C, considering the effects of outlier handling and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). Experiments were conducted on a clinical dataset comprising 615 samples across five diagnostic classes, with four testing scenarios based on combinations of SMOTE application and outlier handling. The results indicate that class balancing using SMOTE consistently improved macro-averaged evaluation metrics, particularly Macro Recall and Macro F1-Score, reflecting the models' enhanced ability to recognize minority classes more evenly. In contrast, outlier handling using the Interquartile Range (IQR) method did not consistently enhance performance across all models. The configuration of Random Forest with SMOTE applied prior to outlier handling achieved the best overall performance, with a Macro F1-Score of 0.8086 and a Macro AUC of 0.9710. These findings emphasize that data balancing plays a crucial role in improving classification performance on clinical datasets with imbalanced class distributions.

REFERENCES

- [1] H. M. Farghaly, M. Y. Shams, and T. A. Abd El Hafeez, "Hepatitis C virus prediction based on a machine learning framework: a real-world case study," *Knowl. Inf. Syst.*, vol. 65, no. 6, pp. 2595–2617, 2023.
- [2] S. K. J. Jang and M. P. Lee, "Development and validation of a machine-learning screening algorithm to improve hepatitis C detection and resource allocation," *Open Forum Infect. Dis.*, vol. 12, no. 8, p. ofaf496, 2025, doi: DOI: 10.1093/ofid/ofaf496.
- [3] Abdul-Mohsen G. and Alhejaily, "Artificial intelligence in healthcare," *Biomed. Rep.*, vol. 22, no. 1, p. 11, 2024, doi: DOI: 10.3892/br.2024.1889.

- [4] E. C. O. P. Rajpurkar, O. Chen, and E. T. Banerjee, "AI in health and medicine," *Nat. Med.*, vol. 28, pp. 31–38, 2021, doi: DOI : 10.1038/s41591-021-01614-0.
- [5] A. M. Edeh, "Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease and fibrosis stage," *J. Med. Syst.*, vol. 46, pp. 1–12, 2022, doi: DOI : 10.1007/s10916-022-01800-x.
- [6] A. M. Alam and et al., "Handling imbalanced medical datasets: review of a decade of progress," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2101–2140, 2024, doi: DOI : 10.1007/s10462-024-10884-2.
- [7] Y. X. Li, S. Zhang, and Liu, "Recent advances in SMOTE-based oversampling for imbalanced classification: a survey," *Inf. Process. Manag.*, vol. 60, p. 102789, 2023, doi: DOI : 10.1016/j.ipm.2023.102789.
- [8] X. Wang and Y. Zhao, "Evaluation metrics for imbalanced classification in medical applications: a practical guideline," *J. Biomed. Inform.*, vol. 128, p. 104000, 2022, doi: DOI : 10.1016/j.jbi.2022.104000.
- [9] J. Zhang, "Advancements of outlier detection: A survey and medical applications," *Electronics / MDPI (survey 2023)*, vol. 13, no. 16, p. 3339, 2023, doi: DOI : 10.3390/electronics13163339.
- [10] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 53, pp. 1275–1308, 2020.
- [11] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques for medical data analysis," *IEEE Access*, vol. 9, pp. 135889–135903, 2021.
- [12] M. L. Wallace, "Use and misuse of random forest variable importance metrics in medicine," *BMC Med. Res. Methodol.*, vol. 23, pp. 1–12, 2023, doi: DOI : 10.1186/s12874-023-01965-x.
- [13] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests in healthcare data classification," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 4, p. e1407, 2021.
- [14] Harsha Nori, Rich Caruana, J. Zhiqi Bu, udy H. Shen, and Janardhan Kulkarni, "Accuracy, interpretability, and differential privacy via Explainable Boosting (DP-EBM)," *PMLR (ICML proceedings), Proceedings of ICML 2021 (PMLR 139)*, 2021.
- [15] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretable machine learning for healthcare," *Commun. ACM*, vol. 64, no. 10, pp. 46–55, 2021.
- [16] R. Kablan, "Evaluation of stacked ensemble model performance to analyze clinical data," *Information Fusion*, 2023, doi: DOI : PubMed ID 37172507 (2023).
- [17] N. Das, "Hepatitis prediction with an explainable machine learning pipeline (includes stacking and SMOTE)," *PLoS One*, 2025, doi: DOI : 10.1371/journal.pone.0319078.
- [18] S. Zhou and Y. Zhang, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Pattern Recognit. Lett.*, vol. 182, pp. 12–24, 2023, doi: DOI : 10.1016/j.patrec.2023.10.012.
- [19] R. Ningsih, A. U. Dullah, and M. A. Muslim, "Hepatitis C classification using neural networks and SMOTE oversampling," *Jurnal Nasional Riset Aplikasi dan Teknik Informatika*, vol. 6, no. 2, pp. 45–54, 2024.
- [20] M. S. Alam, M. M. Rahman, and M. A. Hossain, "A review of machine learning algorithms for hepatitis diagnosis," *Indonesian Journal of Computer Science*, vol. 13, no. 3, pp. 721–735, 2024.
- [21] Jianchang Hu and Silke Szymczak, "A review on longitudinal data analysis with random forest," *Brief. Bioinform.*, vol. 24, no. 2, p. bbad002, 2023, doi: DOI : 10.1093/bib/bbad002.
- [22] Ricardo Gonzalez and et al. Ashirbani Saha, "Seeing the random forest through the decision trees — Supporting learning health systems from histopathology with machine learning models: Challenges and opportunities," *J. Pathol. Inform.*, vol. 15, p. 100347, 2024, doi: DOI : 10.1016/j.jpi.2023.100347.
- [23] Simon P. Couch and Max Kuhn, "Stacks: Stacked ensemble modeling with tidy data principles," *Journal of Open Source Software (JOSS)*, vol. 7, p. 4471, 2022, doi: DOI : 10.21105/joss.04471.
- [24] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, p. 160, 2021.
- [25] H. Khan, M. T. Rasheed, S. Zhang, X. Wang, and H. Liu, "Empirical study of outlier impact in classification context," *Expert Syst. Appl.*, vol. 256, p. 124953, Dec. 2024, doi: 10.1016/j.eswa.2024.124953.
- [26] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, Sep. 2024, doi: 10.1007/s10462-024-10884-2.
- [27] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Health*, vol. 6, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.