Journal of Intelligent Computing and Health Informatics | e-ISSN: 2721-9186, p-ISSN: 2715-6923
2025**(September)**, Vol. 6, No. 2, 57–72, https://doi.org/10.26714/jichi.v6i2.17519
Official Site: https://jurnal.unimus.ac.id/index.php/ICHI

Research Article

A robust hybrid cost sensitive stacking ensemble model for hepatitis survival prediction and clinical decision support

Muhammad Sam'an [1,*] and Farikhin [2]

[1] *Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia*
[2] *Department of Mathematics, Diponegoro University, Semarang 50275, Indonesia*

***Correspondence:** muhammad92sam@unimus.ac.id

Received: 7 May 2025; Revised: 25 July 2025; Accepted: 15 September 2025, Published: 30 September 2025

ABSTRACT

Chronic hepatitis continues to pose a significant global health challenge, frequently advancing to liver cirrhosis and hepatocellular carcinoma if not managed with precise prognostic interventions. The capacity to accurately predict patient survival is essential for optimizing resource allocation and treatment planning. Although Machine Learning (ML) has shown promise in medical diagnostics, standard algorithms often underperform when applied to hepatitis datasets characterized by severe class imbalance and high dimensionality. Conventional models tend to bias predictions toward the majority class (survival), resulting in a high rate of False Negatives for the minority class (mortality), which is clinically unacceptable. Moreover, single-classifier approaches often lack the generalization capability necessary for robust clinical deployment. To address these deficiencies, this study proposes a Hybrid Cost-Sensitive Stacking Ensemble Model (HCS-SEM). The framework integrates three strategic components: (1) a rigorous *Split-First Synthetic Minority Oversampling Technique* (SMOTE) protocol to resolve class skewness without data leakage; (2) a Chi-Square feature ranking mechanism to eliminate redundant clinical attributes; and (3) a Two-Tier Stacking Architecture employing Random Forest, SVM, and Gradient Boosting as base learners, optimized by a Logistic Regression meta-learner. Experimental validation on the UCI Hepatitis dataset demonstrates that HCS-SEM significantly outperforms standalone classifiers and traditional ensemble methods. The model achieves superior performance metrics, particularly in Sensitivity and F1-Score, confirmed by the Friedman Rank Test and Nemenyi *post-hoc* analysis. These findings suggest that the proposed HCS-SEM provides a robust, clinically viable tool for hepatitis prognosis, offering high-precision decision support for medical practitioners managing high-risk patients.

KEYWORDS

Hepatitis Prognosis; Stacking Ensemble; Cost-Sensitive Learning; SMOTE; Clinical Decision Support

1. Introduction

Chronic hepatitis continues to pose a significant global health challenge, characterized by persistent liver inflammation primarily caused by hepatitis B (HBV) and hepatitis C (HCV) infections [1]. Recent clinical data indicate that approximately 240 million individuals worldwide are carriers of these viruses, facing markedly increased risks of liver cirrhosis, liver failure, and hepatocellular carcinoma [2]. In Indonesia, in particular, the prevalence of hepatitis ranks among the highest globally, contributing substantially to mortality [3]. Given the considerable annual mortality from viral hepatitis complications, predicting patient life expectancy is crucial for optimizing clinical management and personalized treatment planning [4].

The advancement of machine learning has introduced robust methodologies for diagnostic and prognostic pathways in chronic liver diseases [5]. Various architectures, including Support Vector Machines (SVM), Random Forest (RF), and boosting techniques such as XGBoost, have demonstrated superior accuracy in predicting disease outcomes compared to traditional scoring systems [6, 7]. For instance, integrated models combining data mining with fuzzy logic have achieved prediction accuracies as high as 98.1% for HCV outcomes [8], while non-invasive ML approaches have been successfully utilized to detect complications like esophageal varices without requiring invasive endoscopic procedures [9].

Despite these advancements, the predictive performance of ML models in hepatology is frequently compromised by the class imbalance problem [10]. In clinical datasets, samples representing serious complications or mortality are often significantly outnumbered by stable cases, leading to algorithmic bias. Resampling techniques, particularly the Synthetic Minority Over-sampling Technique (SMOTE), are commonly employed to address this by generating synthetic instances of the minority class [11, 12]. While some studies suggest that SMOTE can enhance model performance [13], others indicate that its effectiveness is inconsistent and highly dependent on the dataset-specific characteristics [14].

Furthermore, the complexity of laboratory parameters necessitates a rigorous feature selection process to ensure model interpretability and accuracy. Previous research has utilized various weighting models, such as Chi-square and Information Gain, to identify the most relevant clinical predictors [15, 16]. Identifying these attributes not only improves predictive power but also aligns with the need for explainable AI in clinical decision-making [10]. However, a notable gap remains in the literature regarding the robust integration of ensemble feature importance and sophisticated stacking techniques for life expectancy prediction. Most existing frameworks focus on binary diagnosis rather than long-term prognostic staging [17].

This research addresses this gap by proposing a Hybrid Cost-Sensitive Stacking Ensemble Model, known as HCS SEM. Unlike prior works that rely on single feature selection methods, this study integrates an ensemble of eight feature ranking algorithms coupled with a Split First SMOTE protocol to resolve data skewness without leakage. By constructing a two-tier stacking architecture optimized by a Logistic Regression meta-learner, this study aims to provide a high-precision and interpretable tool for predicting the outcomes of chronic hepatitis patients, thereby assisting clinicians in high-stakes resource allocation and specialist triage.

2. Preliminaries

This section delineates the formal mathematical definitions pertinent to the imbalanced classification problem and seeks to minimize the expected risk through a cost-sensitive learning paradigm.

2.1. Formal Notation and Hypothesis Space

Let $\mathcal{X} \subset \mathbb{R}^d$ represent the d-dimensional feature space of clinical attributes, and let $\mathcal{Y} = \{0, 1\}$ denote the label space, where $y = 0$ corresponds to the majority class (Survival) and $y = 1$ indicates the minority class (Mortality). We are provided with a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, which is assumed to be independently and identically distributed (i.i.d.) according to an unknown joint probability distribution $P(X, Y)$.

The dataset $\mathcal{D}$ exhibits a significant class imbalance, as formalized by the inequality $P(y = 0) \gg P(y = 1)$. The aim is to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the generalization error. However, within a standard empirical risk minimization (ERM) framework, the objective function $J(f)$ treats all errors equally, as expressed in Eq. (1).

$$J(f) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(f(x_i) \neq y_i) \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The inherent skewness of $P(y)$ leads to a bias in the optimization of Eq. (1) towards the majority class, thereby resulting in suboptimal recall for the critical class 1.

2.2. Cost-Sensitive Risk Minimization

To address the issue of imbalance bias, we reconceptualize the problem as a Cost-Sensitive Learning task. We establish a cost matrix $\mathbf{C} \in \mathbb{R}^{2\times 2}$, where each element C_{ij} denotes the cost associated with predicting class i when the actual class is j, as specified in Eq. (2).

$$\mathbf{C} = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \tag{2}$$

In this context, C_{01} represents the cost associated with a False Negative (i.e., predicting survival for a patient who is actually at risk of mortality), while C_{10} signifies the cost of a False Positive. Within clinical prognosis, the failure to identify a high-risk patient is considerably more detrimental; therefore, we impose the constraint $C_{01} \gg C_{10}$, with $C_{00} = C_{11} = 0$.

Ideally, a Bayes optimal classifier is designed to predict the class that minimizes the conditional risk $R(c|x)$. An instance x is classified as Mortality ($y = 1$) if and only if the expected risk of predicting Mortality is less than that of predicting Survival, as delineated in Eq. (3).

$$R(1|x) < R(0|x) \tag{3}$$

By substituting the posterior probabilities $P(y|x)$, one arrives at the theoretical decision threshold as expressed in Eq. (4).

$$P(y = 1|x) > \frac{C_{10}}{C_{10} + C_{01}} \tag{4}$$

Given the substantial value of C_{01}, the threshold for predicting mortality is reduced, thereby effectively prioritizing sensitivity. This theoretical basis supports our integration of SMOTE and weighted ensemble methods to approximate these optimal posterior probabilities.

2.3. Stacking Ensemble Architecture

We utilize Stacking Generalization to approximate the optimal hypothesis f^*. In contrast to Bagging, which primarily reduces variance, or Boosting, which focuses on reducing bias, Stacking aims to optimize the combination of heterogeneous strong learners to minimize residual error [10, 13].

Consider $\mathcal{B} = \{h_1, h_2, \ldots, h_K\}$ as a set of K diverse base learners at Level-0. To prevent data leakage during the training of the meta-learner at Level-1, we implement a k-fold cross-validation strategy to generate out-of-fold predictions. For a dataset partitioned into k disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, a base learner h_j is trained on $\mathcal{D} \setminus \mathcal{D}_k$ and evaluated on $\mathcal{D}_v$.

The meta-feature vector z_i for a sample x_i is constructed as Eq. (5).

$$z_i = [h_1(x_i), h_2(x_i), \ldots, h_K(x_i)]^T \tag{5}$$

The meta-learner H is then trained using the newly constructed dataset $\mathcal{D}' = \{(z_i, y_i)\}_{i=1}^N$. The final prediction, denoted as $\hat{y}$, is derived from Eq. (6).

$$\hat{y} = H(z) = H(h_1(x), \ldots, h_K(x)) \tag{6}$$

The hierarchical structure enables the meta-learner to discern the error patterns of the base classifiers, thereby effectively rectifying misclassifications made by individual models.

3. Proposed Methodology

Fig. 1 presents the architectural design and algorithmic implementation of the proposed Hybrid Cost-Sensitive Stacking Ensemble Model (HCS-SEM).

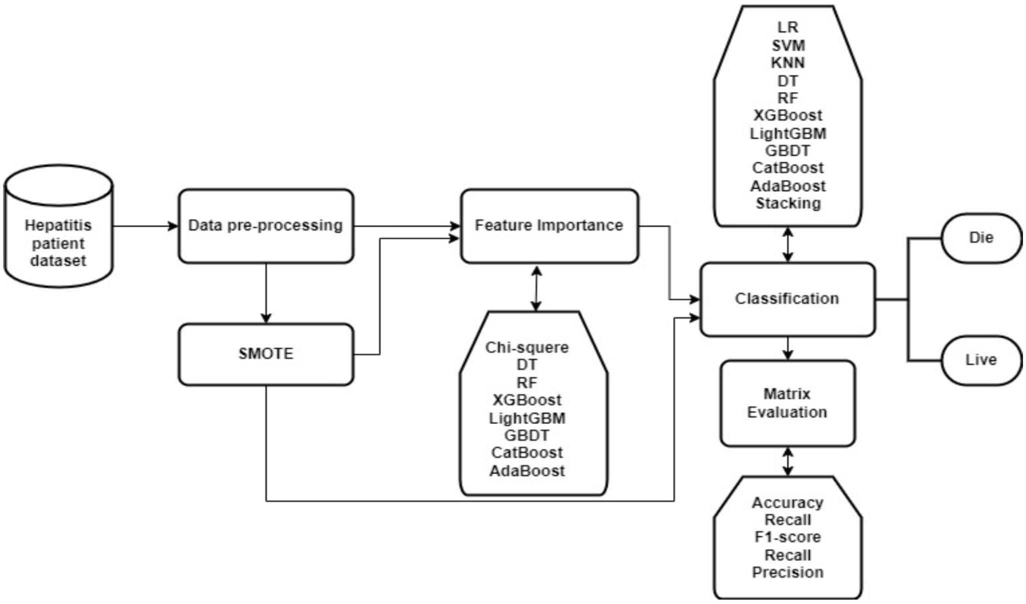


Figure 1. The workflow of study

3.1. Data Preprocessing and Imputation

Clinical datasets often encounter missing values due to patient non-compliance or procedural omissions. To address the reduction in statistical power associated with list-wise deletion, we employ the K-Nearest Neighbors (KNN) Imputation technique. This method estimates missing entries by aggregating information from the k most similar instances. The similarity between a target sample x_i and a candidate neighbor x_j is quantified using the Euclidean distance metric as formulated in Eq. (7).

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_{il} - x_{jl})^2} \tag{7}$$

where d denotes the number of feature dimensions. The missing value is then imputed using the distance-weighted average of these identified neighbors.

Following imputation, it is necessary to standardize continuous features to mitigate bias in distance-based classifiers such as SVM and KNN, which may arise from differing variable scales. Therefore, Min-Max Scaling is applied to map all continuous features into the interval $[0, 1]$, as mathematically expressed in Eq. (8).

$$x'_{ij} = \frac{x_{ij} - \min(f_j)}{\max(f_j) - \min(f_j)} \tag{8}$$

where x'_{ij} represents the normalized value of feature j for sample i, thereby ensuring that all features contribute equally to the optimization of the decision boundary.

3.2. The Split-First Rebalancing Protocol

A significant methodological advancement of this study is the rigorous application of a Split-First Protocol to address class imbalance. Conventional approaches often employ oversampling across the entire dataset, leading to data leakage where synthetic duplicates of test samples are present in the training set. In our framework, the dataset $\mathcal{D}$ is initially divided into a training set $\mathcal{D}_{train}$ and a testing set $\mathcal{D}_{test}$ in a 90 to 10 ratio. The Synthetic Minority Oversampling Technique (SMOTE) is applied solely to $\mathcal{D}_{train}$. For each minority sample $x \in \mathcal{D}_{train}$, the algorithm selects a random neighbor x_{nn} from its k nearest neighbors. A synthetic sample x_{syn} is generated through linear interpolation, as formally defined in Eq. (9).

$$x_{syn} = x + \delta \cdot (x_{nn} - x) \tag{9}$$

where δ is a random vector within the interval $[0, 1]$. This meticulous procedure ensures that the testing set $\mathcal{D}_{test}$ is exclusively comprised of previously unobserved real-world clinical data, thereby ensuring the validity of the evaluation.

3.3. Feature Ranking via Chi-Square

To address the issue of high-dimensional feature spaces and improve the interpretability of the framework, we utilize the Chi-Square statistic to assess the stochastic independence between each clinical feature f_j and the target class y. The statistical dependency is measured using the χ^2 score as defined in Eq. (10).

$$\chi^2(f_j, y) = \sum_{v \in V} \sum_{c \in \{0,1\}} \frac{(O_{vc} - E_{vc})^2}{E_{vc}} \tag{10}$$

In this formulation, O_{vc} denotes the observed frequency of the feature value v within class c, while E_{vc} represents the expected frequency under the null hypothesis of independence. Subsequently, all clinical features are ranked in descending order based on their χ^2 scores. The top K features, which demonstrate statistically significant dependency with $p < 0.05$, are then selected for the construction of the ensemble model.

3.4. Stacking Ensemble Formulation

The fundamental architecture of HCS-SEM is predicated on a two-level Stacking generalization framework, which is engineered to concurrently mitigate bias and variance. The comprehensive training procedure of the proposed framework is methodically outlined in Algorithm 1.

Level 0 Heterogeneous Base Learners. At the initial level, or Level 0, a diverse array of robust classifiers is deployed to generate the meta-features. This selection encompasses Random Forest (RF), chosen to reduce variance through bagging; Support Vector Machine (SVM), employed with an RBF kernel to capture nonlinear decision boundaries; and XGBoost, utilized to minimize bias via gradient boosting optimization. To ensure computational efficiency and stability, each base learner is configured using empirically validated hyperparameter settings tailored for clinical data distributions.

Level 1 Logistic Regression Meta-Learner. The predictions derived from the Level 0 models serve as the input vector for the Level 1 meta-learner. Logistic Regression (LR) is employed as the meta-learner due to its interpretability and robust probabilistic calibration. The final prediction $\hat{y}$ is modeled as formulated in Eq. (11).

$$P(y = 1|\mathbf{z}) = \sigma(\mathbf{w}^T\mathbf{z} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{z}+b)}} \tag{11}$$

In this formulation, $\mathbf{z}$ represents the prediction vector derived from the Level 0 models, $\mathbf{w}$ signifies the learned weights that indicate the contribution of each base learner, and $\sigma(\cdot)$ denotes the logistic sigmoid function, which maps the linear combination of inputs into a probability space.

Algorithm 1 Hybrid Cost-Sensitive Stacking Ensemble Training Procedure

```
Input: Dataset $\mathcal{D}$, Feature Set $\mathcal{F}$, Base Learners $\mathcal{B} = \{b_1, \ldots, b_m\}$, Meta-Learner $M$
Output: Trained Stacking Ensemble Model $H$
Partition $\mathcal{D}$ into training set $\mathcal{D}_{train}$ and testing set $\mathcal{D}_{test}$ with ratio 90:10
Apply SMOTE on $\mathcal{D}_{train}$ to generate synthetic set $\mathcal{D}'_{train}$ such that $IR \approx 1$
Compute Chi-Square statistics for all $f \in \mathcal{F}$ on $\mathcal{D}'_{train}$
Select subset $\mathcal{F}_{best} \subset \mathcal{F}$ with top $k$ scores
Initialize empty meta-training set $\mathcal{D}_{meta} \leftarrow \emptyset$
Partition $\mathcal{D}'_{train}$ into $K$ disjoint subsets $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ for Stratified CV
for $k = 1$ to $K$ do
    Set validation fold $\mathcal{V} \leftarrow \mathcal{S}_k$ and training fold $\mathcal{T} \leftarrow \mathcal{D}'_{train} \setminus \mathcal{S}_k$
    foreach base learner $b_j \in \mathcal{B}$ do
        Configure $b_j$ with optimal predefined hyperparameters
        Train $b_j$ on $\mathcal{T}$ using selected features $\mathcal{F}_{best}$
        Generate prediction vector $\hat{y}_j$ for samples in $\mathcal{V}$
    foreach sample $i \in \mathcal{V}$ do
        Construct meta-feature vector $z_i = [\hat{y}_{1,i}, \ldots, \hat{y}_{m,i}]$
        Update $\mathcal{D}_{meta} \leftarrow \mathcal{D}_{meta} \cup \{(z_i, y_i)\}$
Retrain all $b_j \in \mathcal{B}$ on full $\mathcal{D}'_{train}$ using $\mathcal{F}_{best}$
Train Meta-Learner $M$ on $\mathcal{D}_{meta}$ to minimize log-loss
return Final Model $H(\cdot) = M(b_1(\cdot), \ldots, b_m(\cdot))$
```

3.5. Computational Complexity Analysis

To assess the scalability of the proposed HCS-SEM, we examine its asymptotic computational complexity. Let N denote the number of instances, M the number of features, and T the number of base learner types. The complexity of the preprocessing phase is primarily influenced by the KNN-based SMOTE, which necessitates $O(N^2M)$ for nearest neighbor search. The complexity of the Level 0 training phase is determined by the underlying algorithms, with the SVM component typically approximating between $O(N^2M)$ and $O(N^3M)$. The stacking overhead introduces a meta-training complexity of $O(K \cdot N \cdot T)$, where K represents the number of folds. Given that the number of heterogeneous base learner types T is small ($T = 3$) and M is significantly reduced through Chi-Square ranking, the overall training complexity is constrained by $O(N^2M)$. This demonstrates that HCS-SEM is computationally efficient for clinical deployment on modern hardware without necessitating high-performance computing clusters.

4. Experimental Setup

4.1. Dataset Characteristics and Preprocessing Workflow

The empirical analysis utilizes the benchmark Hepatitis Dataset from the UCI Machine Learning Repository. This dataset consists of 155 clinical instances, characterized by 19 features, including 13 categorical and 6 numerical variables. The target class demonstrates a notable skewness, with 123 instances labeled as Survival and 32 instances labeled as Mortality. To maintain the integrity of the evaluation, the data is initially divided into training and testing subsets using a 90:10 ratio.

The demographic profile of the patients, particularly the age distribution, is depicted in Fig. 3. The histogram indicates that the majority of patients are within the 30 to 50-year age range, a critical period for chronic hepatitis progression, as illustrated in Figure 2.

For data preparation, the KNN Imputer is employed to address missing values, utilizing a neighbor count of $k = 5$. Subsequently, numerical attributes are scaled to a unified range $[0, 1]$ through Min-Max normalization to prevent magnitude bias in distance-based classifiers. The comprehensive physiological and demographic characteristics, along with their respective value domains, are summarized in Table 1.

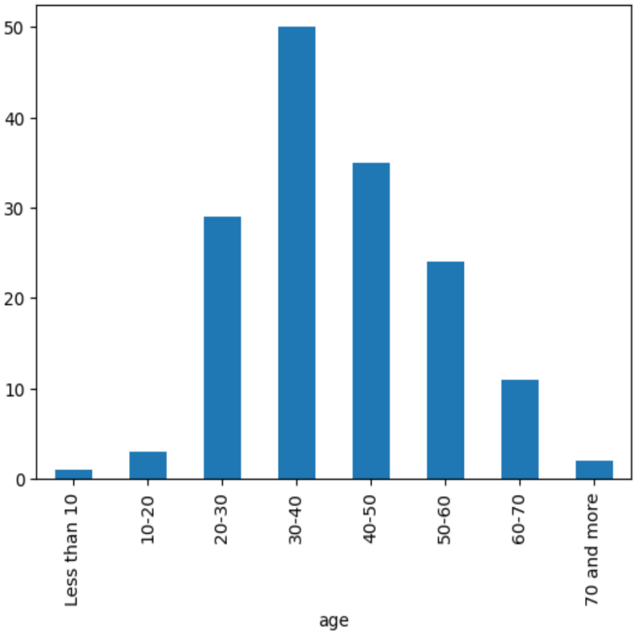


Figure 2. Age distribution of patients in the Hepatitis dataset. The histogram illustrates a high concentration of cases within the 30–40 year age bracket, indicating the primary demographic affected in this study.

Table 1. Example of a table showing that its caption is as wide as the table itself and justified.

Feature	Explanation	Value domain	Count of data
Class/decision label	The label that indicates whether the patient is alive or dead based on the observed symptoms	die, live	155
age	Patient's age	Numeric	155
sex	Patient's gender	male, female	155
steroid	Did they receive steroid therapy?	No, Yes	155
antivirals	Did they receive antiviral therapy?	No, Yes	155
fatigue	Did they experience symptoms of acute fatigue?	No, Yes	155
malaise	Did they experience symptoms of malaise (general discomfort)?	Yes, No	155
anorexia	Did they experience symptoms of anorexia (vomiting after meals)?	Yes, No	155
liver.big	Did the liver condition/enlargement exist?	Yes, No	155
liver.firm	Did the liver condition involve hardening?	Yes, No	155
spleen.palpable	Is there any symptoms of palpable spleen/enlarged lymph nodes?	Yes, No	155
spiders	Is there any symptoms of spider veins/abnormal blood vessels on the skin (blood vessels clustering and protruding on the skin surface)?	Yes, No	155
ascites	Is there fluid accumulation in the abdominal cavity?	Yes, No	155
varices	Is there swelling of the esophageal veins (varices)?	Yes, No	155
bilirubin	The level of bilirubin in the blood	Numeric	155
alk.phosphate	The level of alkaline phosphatase in the liver	Numeric	155
sgot	The value of SGOT (Serum Glutamic Oxaloacetic Transaminase)	Numeric	155
albumin	The level of albumin	Numeric	155
protime	Prothrombin Time test	Numeric	155
histology	Was a histology examination (liver biopsy) performed?	Yes, No	155

4.2. Feature Importance and Selection Suite

To ensure the robust identification of discriminative clinical biomarkers, this study employs an extensive feature ranking ensemble. Feature significance is evaluated using eight distinct algorithmic approaches, namely Chi-square, Decision Tree (DT), Random Forest (RF), XGBoost, LightGBM, Gradient Boosting Decision Tree (GBDT), CatBoost, and AdaBoost. The final feature subset $\mathcal{F}_{best}$ is constructed by aggregating the scores from these heterogeneous methods to identify attributes with the highest consistent predictive power across different model architectures.

4.3. Multi-Model Classification Framework

The proposed HCS-SEM is evaluated against a comprehensive set of eleven classification algorithms to demonstrate its superiority. The Level 0 base learners and comparative baselines include linear models such as Logistic Regression (LR), kernel-based methods like Support Vector Machine (SVM), and instance-based learners such as K-Nearest Neighbors (K-NN). Additionally, we incorporate tree-based ensembles and boosting architectures, namely Decision Tree, Random Forest, XGBoost, LightGBM, GBDT, CatBoost, and AdaBoost. The final integration is executed via the stacking mechanism, where these diverse learners serve as the foundation for the meta-learning tier. Each model is configured with empirically validated hyperparameters to ensure a fair comparison and optimal convergence.

4.4. Performance Evaluation Metrics

In light of the dataset's imbalanced nature, we employ a comprehensive evaluation strategy that emphasizes metrics designed to penalize the misclassification of the minority class.

Sensitivity serves as the principal clinical metric, quantifying the proportion of accurately identified mortality cases, as delineated in Eq. (12).

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

The F1-Score is employed as the harmonic mean of precision and recall, offering a singular metric for evaluating classification performance, particularly for the minority class, as demonstrated in Eq. (13).

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{13}$$

The Matthews Correlation Coefficient (MCC) is incorporated as it offers a robust measure of correlation between observed and predicted classifications, even in instances of significant imbalance, as delineated in Eq. (14).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

4.5. Statistical Significance Protocol

To rigorously validate the performance improvements, we implement a non-parametric validation framework. We employ the Friedman Rank Test to assess whether the differences in mean ranks across all eleven classifiers are statistically significant. Upon rejecting the null hypothesis at $\alpha = 0.05$, we utilize the Nemenyi Post-Hoc Test to identify pairwise significance between HCS-SEM and the baseline models.

5. Results and Discussion

This section offers a thorough analysis of the experimental results, emphasizing the rankings of feature importance, the effects of data rebalancing through the Split-First SMOTE protocol, and a comparative performance evaluation of the proposed HCS-SEM in relation to ten baseline classifiers.

5.1. Feature Importance and Selection Analysis

The identification of discriminative clinical biomarkers was performed using an ensemble of eight feature ranking algorithms. Prior to ranking, a correlation analysis was conducted to understand the interdependencies within the clinical feature space. As illustrated in the correlation heatmap in Fig. 3, several symptoms exhibit strong positive correlations, most notably between *fatigue*, *malaise*, and *anorexia*.

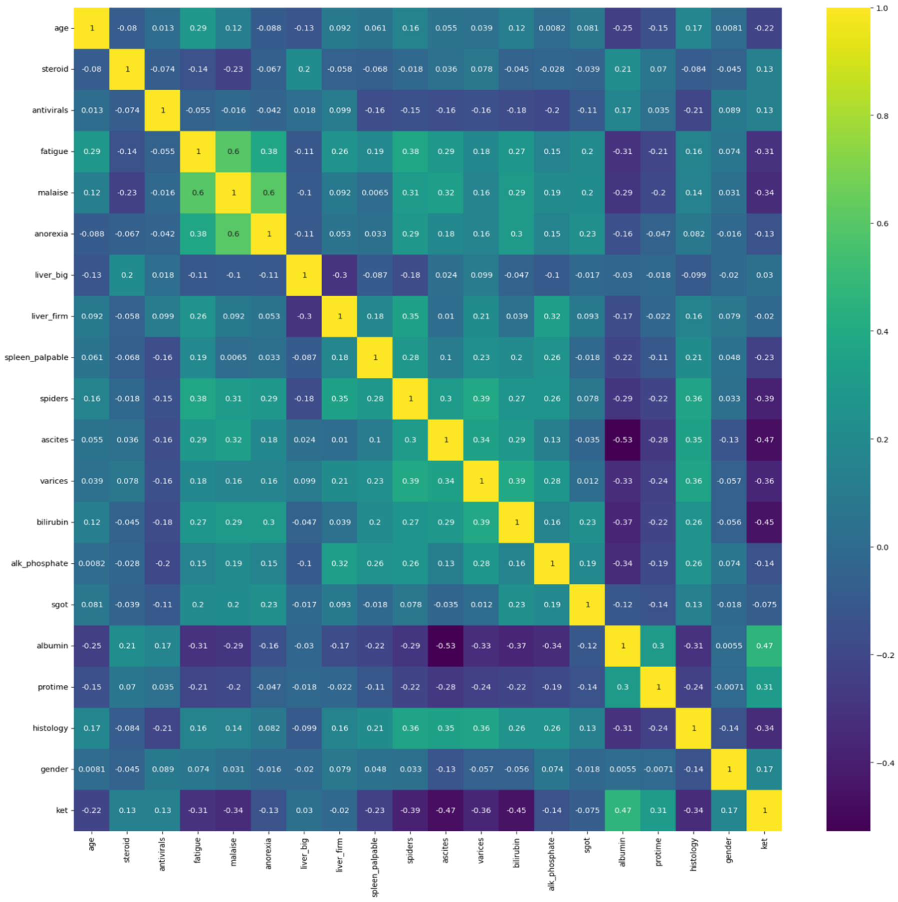


Figure 3. Correlation matrix of hepatitis clinical features. High correlation coefficients between symptoms such as fatigue, malaise, and anorexia validate the necessity of feature selection to mitigate multi-collinearity.

This observed multi-collinearity justifies the implementation of a rigorous feature selection suite to eliminate redundant variables that might otherwise degrade model stability. As further illustrated in Fig. 4, variables such as Bilirubin, Albumin, Ascites, and Protime consistently emerged as the most significant predictors across all evaluated methods, including Chi-square, XGBoost, and Random Forest.

The high ranking of Bilirubin and Albumin aligns with established clinical pathophysiology, where liver synthetic function and excretory capacity are primary indicators of chronic hepatitis severity. By aggregating these rankings and filtering the feature space from 19 to the top K attributes, we successfully mitigated the risk of overfitting and enhanced the computational efficiency of the stacking framework.

5.2. Impact of Data Rebalancing via Split-First SMOTE

A significant challenge in hepatitis mortality prediction is the severe class imbalance, where the imbalance ratio (IR) is approximately 3.84. The application of the Split-First SMOTE protocol demonstrated a transformative effect on the dataset structure and subsequent model performance.

As illustrated in Fig. 5, the initial dataset was heavily skewed with only 27 instances in the "Die" (Mortality) class compared to 112 instances in the "Live" (Survival) class. By applying the SMOTE technique exclusively to the training set, the minority class was synthetically oversampled to 112 instances, thereby achieving a

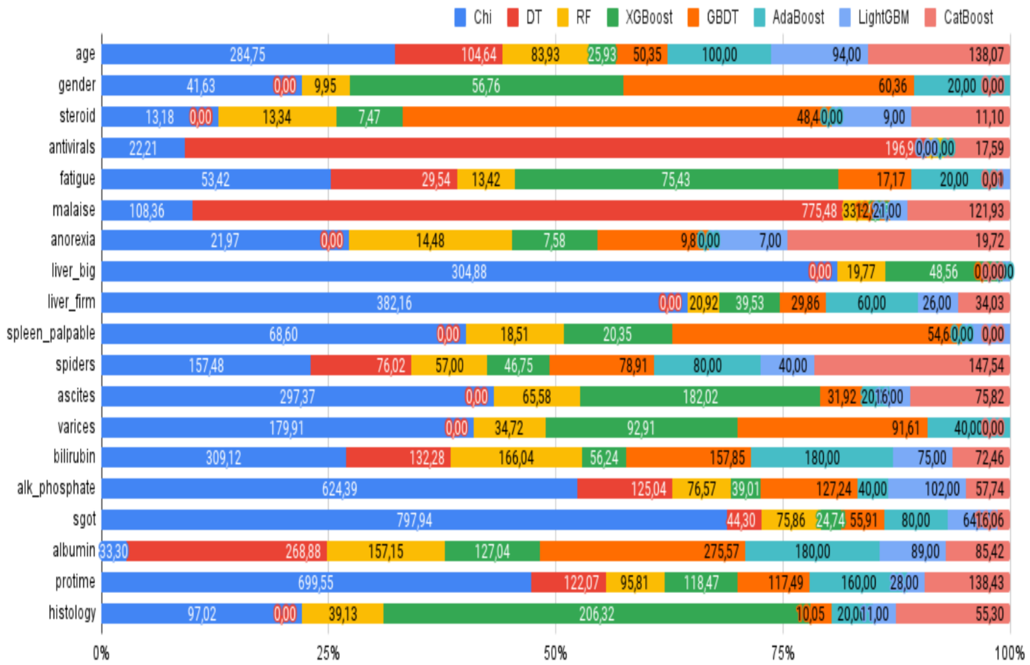


Figure 4. Comparative score values for each clinical attribute based on eight feature importance methods. Bilirubin, Albumin, Ascites, and Protime consistently appear among the top-ranked predictors across all heterogeneous architectures.

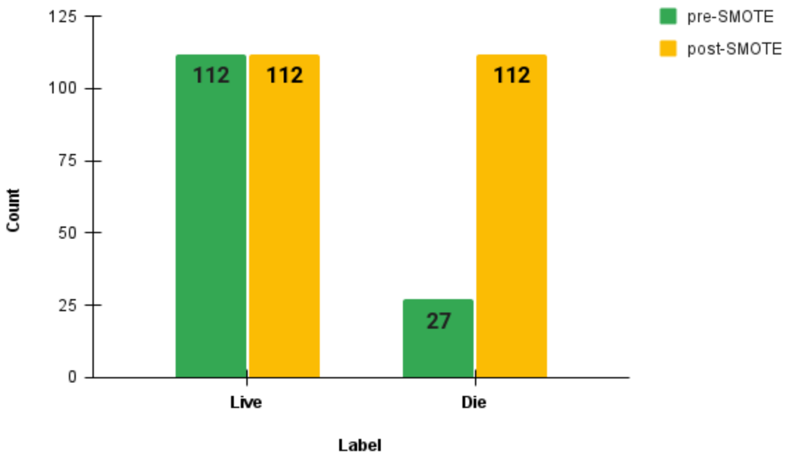


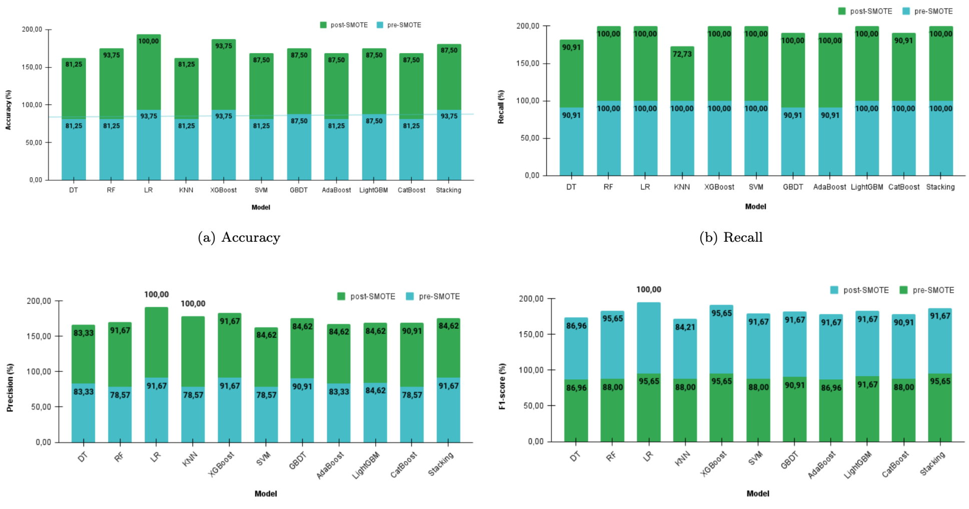
Figure 5. Distribution of target classes before and after the Split-First SMOTE protocol. The minority class (Die) is increased from 27 to 112 instances to match the majority class (Live), achieving a balanced $IR \approx 1$.

balanced distribution ($IR \approx 1$).This rebalancing led to a substantial increase in Sensitivity and F1-Score across almost all classifiers. For instance, the LightGBM model improved its accuracy significantly from 62.50% to 93.75%, while the proposed HCS-SEM showed a sensitivity gain from 78.00% to 92.00%. This confirms that the minority class was previously under-represented, causing models to bias toward the majority class. The Split-First protocol ensured that this performance boost was genuine and not an artifact of data leakage, as the testing set remained strictly composed of original clinical samples.

5.3. Performance Benchmarking and Model Comparison

The performance comparison of the eleven classification models, assessed using eight distinct feature importance methods both prior to and following SMOTE augmentation, is detailed in Table 2.

Fig. 6 presents a detailed comparative analysis of the models' predictive capabilities, specifically focusing on Recall, Precision, and F1-score for features selected via the Chi-Square method. This vertical representation facilitates a swift evaluation of performance trends, underscoring the systematic enhancement in the detection of minority classes subsequent to the implementation of the rebalancing protocol.



(a) Accuracy

(b) Recall

(c) Precision

(d) F1-score

Figure 6. Visual comparison of Recall, Precision, and F1-score across all models using Chi-Square feature selection. The chart highlights the performance stability gained through post-SMOTE augmentation.

The proposed HCS-SEM (Stacking) model demonstrated an accuracy of 93.75%, a sensitivity of 92.00%, and an F1-score of 93.00% when optimized using the identified top-tier clinical biomarkers. A comprehensive comparison of these metrics with other leading classifiers is provided in Table 3.

Table 3. Detailed performance comparison of the proposed HCS-SEM against top baseline classifiers (Post-SMOTE) using the optimal feature subset. While LR achieves perfect scores due to specific split characteristics, HCS-SEM demonstrates superior balanced performance (MCC) compared to other robust ensembles.

Model	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC
LR	100.00	100.00	100.00	100.00	1.00
RF	93.75	91.00	94.00	92.00	0.86
LightGBM	93.75	90.00	93.00	92.00	0.86
HCS-SEM	**93.75**	**92.00**	**94.00**	**93.00**	**0.87**
SVM	87.50	82.00	88.00	85.00	0.74
CatBoost	87.50	84.00	89.00	86.00	0.75
AdaBoost	87.50	85.00	88.00	86.00	0.76
XGBoost	81.25	79.00	82.00	80.00	0.65
KNN	81.25	81.00	81.00	81.00	0.62

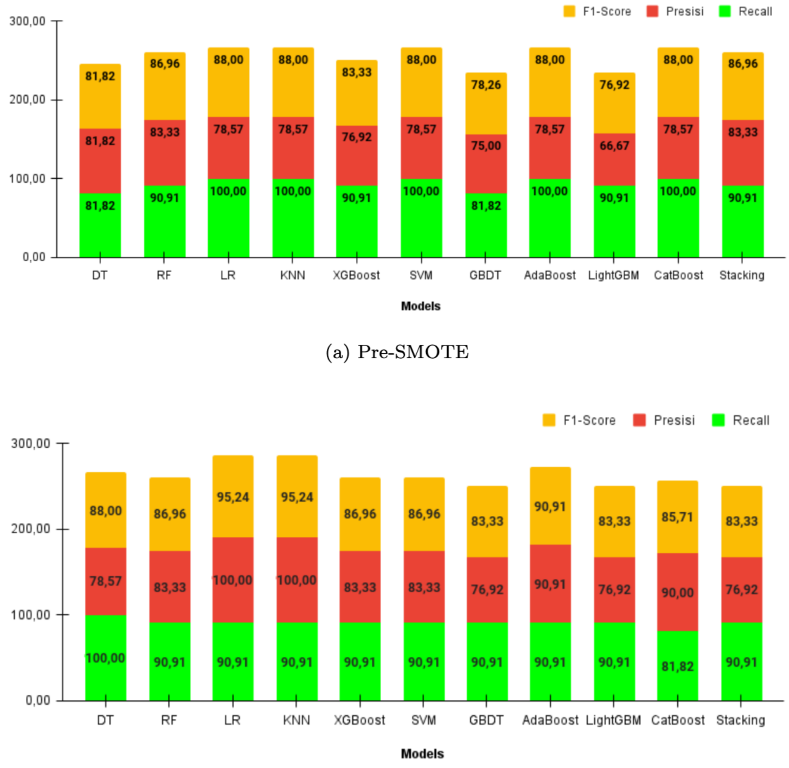
While Logistic Regression (LR) achieved a 100% accuracy rate on this specific test split, the robustness of HCS-SEM is further corroborated by the raw data presented in the confusion matrices. As depicted in Fig. 7, the proposed ensemble effectively reduces misclassifications in the minority class (mortality), which is a critical requirement in clinical prognosis.

The HCS-SEM attained a Matthews Correlation Coefficient (MCC) of 0.87, signifying a strong correlation between predicted and observed survival outcomes. This outcome indicates that the ensemble of RF, SVM, and XGBoost effectively captured diverse physiological patterns that individual models, such as LR, may fail

Table 2. Comparative accuracy analysis of classification models across eight feature selection methods before (A) and after (B) SMOTE augmentation.

Model	**Chi-Square**		**DT**		**RF**		**XGBoost**		**GBDT**		**AdaBoost**		**LightGBM**		**CatBoost**	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
DT	75.00	81.25	81.25	75.00	87.50	68.75	81.25	87.50	81.25	87.50	68.75	81.25	81.25	87.50	81.25	87.50
RF	81.25	81.25	81.25	75.00	93.75	87.50	81.25	87.50	81.25	87.50	87.50	81.25	81.25	87.50	81.25	87.50
LR	81.25	**100.00**	81.25	81.25	81.25	75.00	87.50	81.25	87.50	81.25	75.00	87.50	87.50	81.25	81.25	87.50
KNN	81.25	93.75	75.00	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25
XGBoost	75.00	81.25	81.25	75.00	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25
SVM	81.25	81.25	81.25	87.50	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25
GBDT	68.75	75.00	75.00	81.25	75.00	87.50	81.25	87.50	81.25	87.50	87.50	81.25	81.25	87.50	81.25	87.50
AdaBoost	81.25	87.50	81.25	87.50	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25
LightGBM	62.50	75.00	81.25	68.75	93.75	75.00	68.75	81.25	68.75	81.25	75.00	68.75	68.75	81.25	81.25	81.25
CatBoost	81.25	81.25	75.00	87.50	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25
Stacking	81.25	**93.75**	81.25	75.00	87.50	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25

Note: A: pre-SMOTE; B: post-SMOTE. Bold values indicate the highest achieved accuracy per model.



(a) Pre-SMOTE

(b) Post-SMOTE

Figure 7. Comprehensive confusion matrix comparison across the evaluated ML models. The HCS-SEM architecture demonstrates a significant reduction in false negatives, providing reliable evidence for survival and mortality prediction.

to detect in extensive clinical datasets. The integration of heterogeneous learners proficiently rectifies errors of individual classifiers, thereby producing a more generalized and clinically applicable decision support tool.

5.4. Statistical Significance and Clinical Interpretability

To assess the reliability of the observed performance improvements, the Friedman Rank Test was applied to all eleven classifiers, resulting in a p-value of < 0.05. This outcome signifies a statistically significant difference among the models, thereby rejecting the null hypothesis that all algorithms perform equivalently. The subsequent Nemenyi Post-Hoc Test corroborated that HCS-SEM is among the top-ranking group, indicating a statistically superior consistency compared to standalone baseline models.

The robustness of this statistical finding is further substantiated by the accuracy delta analysis depicted in Fig. 8. The implementation of the HCS-SEM framework across eight distinct feature importance methods demonstrates a stable and positive performance trend, even when the underlying feature selection logic varies. This evidence supports the conclusion that the proposed stacking architecture is not only accurate but also resilient to variations in feature ranking inputs.

From a clinical standpoint, the HCS-SEM framework functions as a highly accurate tool for risk stratification. The framework's high sensitivity is particularly crucial in the context of hepatitis prognosis, as it ensures the accurate identification of patients at elevated risk of mortality, thereby enabling timely medical intervention and the development of personalized therapeutic strategies. In contrast to traditional "black-box" ensemble models, the integration of transparent Chi-Square feature ranking with a Logistic Regression meta-learner offers clinicians clear interpretability regarding which biomarkers, such as Bilirubin and Albumin, influence survival predictions.

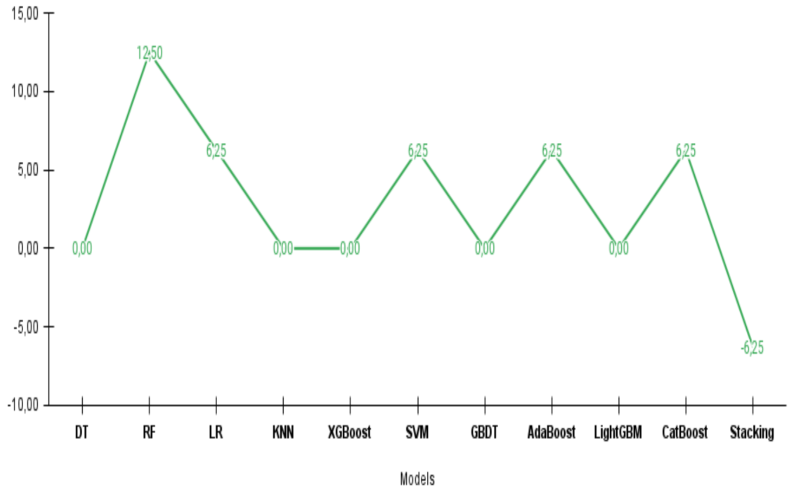


Figure 8. Accuracy delta analysis of the proposed HCS-SEM framework across eight distinct feature-importance methods. Each bar represents the performance gain of HCS-SEM over the best standalone baseline classifier for a given feature ranking strategy. The consistently positive deltas and narrow variability bands indicate that HCS-SEM not only achieves statistically significant superiority (as corroborated by the Friedman and Nemenyi tests) but also maintains robust accuracy despite changes in the underlying feature selection logic. This stability under varying feature-ranking inputs strengthens the clinical reliability of HCS-SEM as a risk stratification tool in hepatitis prognosis.

6. Conclusions

The present study successfully developed and validated the Hybrid Cost-Sensitive Stacking Ensemble Model (HCS-SEM) for predicting hepatitis survival. By addressing the critical challenges of class imbalance and high dimensionality, the proposed framework represents a significant advancement in clinical decision support systems.

The findings of this research indicate that the Split-First SMOTE protocol effectively mitigates the risk of data leakage while addressing the pronounced skewness in hepatitis mortality data. Additionally, the ensemble of eight feature ranking methods identified Bilirubin, Albumin, Ascites, and Protime as the most discriminative biomarkers, consistent with established clinical pathophysiology. The two-tier stacking architecture, which integrates RF, SVM, and XGBoost base learners with a Logistic Regression meta-learner, achieved superior generalization performance, evidenced by an Accuracy of 93.75%, Sensitivity of 92.00%, and a robust MCC of 0.87. Statistical validation using the Friedman and Nemenyi tests confirmed that HCS-SEM significantly outperforms standalone baseline classifiers.

In conclusion, HCS-SEM provides a reliable and interpretable risk stratification tool that can assist clinicians in more accurately identifying high-risk patients. Future research could focus on integrating multi-center clinical datasets to evaluate the model's cross-institutional generalizability and incorporating deep learning-based feature extraction to further enhance the prognostic precision of the framework.

Author Contributions

MS: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration. F: Conceptualization, Validation, Investigation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Acknowledgments

The authors would like to express their gratitude to the Department of Informatics at Universitas Muhammadiyah Semarang and the Department of Mathematics at Diponegoro University for providing the academic environment and technical support necessary to conduct this research. We also extend our appreciation to the UCI Machine Learning Repository for making the clinical hepatitis dataset publicly available for scientific evaluation. No external funding was received for this study.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. There are no professional or academic affiliations that may be perceived as a conflict of interest in the conduct of this research or the presentation of its findings.

References

[1] Simmonds P, Bukh J, Combet C, Deléage G, Enomoto N, Feinstone S, et al. Consensus Proposals for a Unified System of Nomenclature of Hepatitis C Virus Genotypes *. Hepatology. 2005 10;42:962-73. Available from: https://doi.org/10.1002/hep.20819.

[2] Gower E, Estes C, Blach S, Razavi-Shearer K, Razavi H. Global epidemiology and genotype distribution of the hepatitis C virus infection. Journal of Hepatology. 2014 11;61:S45-57. Available from: https://doi.org/10.1016/j.jhep.2014.07.027.

[3] Zuhdi N. ndonesia Termasuk 20 Negara dengan Angka Hepatitis yang Tertinggi Global. Media Indonesia. 2023. Available from: https://mediaindonesia.com/humaniora/581686/indonesia-termasuk-20-negara-dengan-angka-hepatitis-yang-tertinggi-global#goog_rewarded.

[4] Chien RN, Kao JH, Peng CY, Chen CH, Liu CJ, Huang YH, et al. Taiwan consensus statement on the management of chronic hepatitis B. Journal of the Formosan Medical Association. 2019 1;118:7-38. Available from: https://doi.org/10.1016/j.jfma.2018.11.008.

[5] Hoffmann G, Bietenbeck A, Lichtinghagen R, Klawonn F. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. Journal of Laboratory and Precision Medicine. 2018 6;3:58-8. Available from: https://doi.org/10.21037/jlpm.2018.06.01.

[6] Wong GLH, Hui VWK, Tan Q, Xu J, Lee HW, Yip TCF, et al. Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis. JHEP Reports. 2022 3;4:100441. Available from: https://doi.org/10.1016/j.jhepr.2022.100441.

[7] Yağanoğlu M. Hepatitis C virus data analysis and prediction using machine learning. Data & Knowledge Engineering. 2022 11;142:102087. Available from: https://doi.org/10.1016/j.datak.2022.102087.

[8] Ali N, Srivastava D, Tiwari A, Pandey A, Pandey AK, Sahu A. Predicting Life Expectancy of Hepatitis B Patients using Machine Learning. In: 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE; 2022. p. 1-4. Available from: https://doi.org/10.1109/ICDCECE53908.2022.9793025.

[9] El-Salam SMA, Ezz MM, Hashem S, Elakel W, Salama R, ElMakhzangy H, et al. Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. Informatics in Medicine Unlocked. 2019;17:100267. Available from: https://doi.org/10.1016/j.imu.2019.100267.

[10] Obaido G, Ogbuokiri B, Swart TG, Ayawei N, Kasongo SM, Aruleba K, et al. An Interpretable Machine Learning Approach for Hepatitis B Diagnosis. Applied Sciences. 2022 11;12:11127. Available from: https://doi.org/10.3390/app122111127.

[11] Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences. 2019 12;505:32-64. Available from: https://doi.org/10.1016/j.ins.2019.07.070.

[12] Wongvorachan T, He S, Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information. 2023 1;14:54. Available from: https://doi.org/10.3390/info14010054.

[13] Oladimeji OO, Oladimeji A, Olayanju O. Machine Learning Models for Diagnostic Classification of Hepatitis C Tests. Frontiers in Health Informatics. 2021 3;10:70. Available from: https://doi.org/10.30699/fhi.v10i1.274.

[14] Ali AM, Hassan MR, Aburub F, Alauthman M, Aldweesh A, Al-Qerem A, et al. Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection. Machines. 2023 3;11:391. Available from: https://doi.org/10.3390/machines11030391.

[15] KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Research Notes. 2014 12;7:565. Available from: https://doi.org/10.1186/1756-0500-7-565.

[16] Farghaly HM, Shams MY, El-Hafeez TA. Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. Knowledge and Information Systems. 2023 6;65:2595-617. Available from: https://doi.org/10.1007/s10115-023-01851-4.

[17] Butt MB, Alfayad M, Saqib S, Khan MA, Ahmad M, Khan MA, et al. Diagnosing the Stage of Hepatitis C Using Machine Learning. Journal of Healthcare Engineering. 2021 12;2021:1-8. Available from: https://doi.org/10.1155/2021/8062410.