A Machine Learning Framework for Non-invasive Hepatitis C Virus Detection Using Feature Selection

***Abstract*—Early and accurate detection of the Hepatitis C Virus (HCV) is critical for timely treatment, especially in low-resource settings where traditional diagnostic methods are invasive and expensive. A machine learning framework for non-invasive multiclass HCV classification using standard laboratory data is presented in this paper. Class imbalance and high-dimensional, noisy features are two major issues in medical datasets that the framework tackles. We apply a generative AI-based oversampling technique using the Synthetic Data Vault (SDV) to balance the dataset, followed by feature selection methods—Recursive Feature Elimination (RFE), Minimum Redundancy Maximum Relevance (mRMR), and Boruta—to identify optimal feature subsets. Three classifiers—Logistic Regression, Random Forest, and XGBoost—are trained and evaluated using F1-score and ROC-AUC metrics. SHAP-based explainability analysis is also conducted to interpret feature contributions.The proposed framework, particularly with RFE and SDV-based augmentation, achieves high classification accuracy (96.7%) and improved F1-scores for minority classes, demonstrating its effectiveness for robust and interpretable HCV multiclass detection.**

***Index Terms*—Imbalanced classification, Synthetic data generation, SDV, Feature selection, Recursive Feature Elimination (RFE), Logistic Regression, SHAP, Hepatitis C, Multiclass classification, Explainable AI, Medical diagnosis.**

I. Introduction

Hepatitis C virus (HCV) poses a significant global health challenge, affecting an estimated 58 million people worldwide and causing over 290,000 deaths annually due to complications such as liver cirrhosis and hepatocellular carcinoma [1]. In specific regional context, HCV remains an important cause of mortality in Bangladesh. A study conducted near Dhaka found that 0.88% if individuals tested positive for anti-HCV antibodies, with predominance among males aged 17-50 [2]. To handle the current situation, early and accurate detection of HCV is crucial for effective treatment and long-term monitoring. However, traditional procedures- including liver biopsies, elastography, and molecular testing- are often invasive, costly and require specialized infrastructure, making them less accessible in low-resource settings [3].

Recent advancements in machine learning techniques have revolutionized in development of non-invasive, data-driven diagnostic models that use routine laboratory measurements such as serum biochemistry and hematology. However, real-world medical datasets have several challenges for classification tasks. Notably, they are imbalanced, with minority classes (e.g. early stage or asymptomatic HCV) underrepresented, leading to a biased classifier and poor generalization on rare cases. Additionally, the high dimensional features and missing values in medical datasets further compromises the model's performance in classification tasks [4].

To address the issues, we propose a machine learning framework for multiclass HCV classification using non-invasive feature set. Our work focuses on handling the imbalanced dataset, selecting relevant features, and building interpretable and effective multiclass classifiers. We use Synthetic Data Vault (SDV), a generative AI-based oversampling technique that synthesizes realistic minority class samples by learning the joint distribution of data, to handle the class imbalance. Then to reduce the feature set, we apply three feature selection method namely, Recursive Feature Elimination (RFE), Minimum Redundancy Maximum Relevance (mRMR), and Boruta, to identify the most important and compact feature subset that preserve the predicitive power. For classification task, we compare the classification performance of Logistic Regression, Random Forest, and XGBoost, using accuracy, F1-score, and Area under the Receiver Operating Characteristic Curve (AUC-ROC) as evaluation metrics. Finally, we apply SHapley Additive exPlanations (SHAP) to interpret model predictions and highlight the relevence of our selected features contributing to detection of multiple classes in HCV classification.

II. Related Work

Recent works on machine learning based Hepatitis C Virus(HCV) has mainly focused on three key areas: feature selection, data preprocessing and classification strategies.

The comprative analysis on feature selection and without feature selection was studied and it demonstrates significant improvement for less features in binary classification of Hepatitis C patient detection [5]. Building on this, a novel meta-model with pre-clustering approach for feature selection was proposed and it performed well with and accuracy of 94.82% on binary classification [6]. In most of the studies on HCV detection, quality of data significantly impacts the model performance. In [7], the work suggested that median imputation on missing values helped to gain accurate results. An automatic classifier for multiclass probabilities of HCV incidence based on patients' blood attributes. The work handled the class imbalance on the dataset using synthetic minority oversampling technique (SMOTE) and used an ensamble based RF-LR model which gained a high accuracy [8]. A personalized machine learning approach was taken for HCV binary classification by tailoring data preprocessing and hyperparameter tuning that achieved 99% accuracy and 94% recall [9]. Various rule based model based on multiple

Decision Trees was developed but these works don't perform well in case of multiclass classification for HCV detection [10], [11].

Existing literature on HCV detection performs well for binary classification, but significant gaps remain in multiclass accuracy due to lack of handling class imbalance properly, which our personalized modeling approach direcly addresses.

III. Methodology

A. Dataset

The dataset used in this study was obtained from the University of California Irvine Machine Learning Repository (UCI-MLR) [12]. It contains demographic and clinical laboratory data for 615 individuals. It included blood donors and Hepatitis C patients at different stages of the disease: Hepatitis, Fibrosis, and Cirrhosis. The original dataset had a "suspect blood donor" class, which had only seven instances, so for better model performance, this class was excluded from this study. Therefore, the target variable includes four classes: 0 for blood donors, 1 for Hepatitis, 2 for Fibrosis, and 3 for Cirrhosis.

The original dataset contains 14 columns, but the patient ID column was excluded from the analysis, resulting in 13 meaningful features. These include demographic attributes such as Age and Sex, and various laboratory biomarkers like Albumin (ALB), Alkaline Phosphatase (ALP), Alanine Amino-Transferase (ALT), Aspartate Amino-Transferase (AST), Bilirubin (BIL), Choline Esterase (CHE), Cholesterol (CHOL), Creatinine (CREA), Gamma Glutamyl Transferase (GGT), and Total Protein (PROT).

All laboratory features are continuous, while Sex and Category are categorical variables. Age is a numerical demographic variable. Several features contain missing values, which were handled using appropriate imputation techniques during the preprocessing stage.

B. Data Preprocessing

To prepare the dataset for further analysis, One-hot encoding was applied to the Sex column because this allows the model to interpret it numerically. After that, scaling is applied to other numerical features. For scaling, StandardScaler is used. It ensures all features have equal importance during the learning process and scales the overall range between 0 and 1.

For handling missing values, a gender-specific median imputation strategy was used. Specifically, for each lab-related feature with missing entries, the median was calculated separately for male and female patients. This method was chosen because in almost every column, there is a slight difference between the male and female median values.

C. Synthetic Data Generation

To address class imbalance and enrich the dataset, synthetic samples were generated using the TVAESynthesizer from the Synthetic Data Vault (SDV) [13]. A metadata schema was defined to inform the model of each feature's type.

The quality and validity of the synthetic data were assessed using SDV's built-in tools:

- `run_diagnostic()` to identify modeling issues,
- `evaluate_quality()` to assess the statistical similarity between real and synthetic data, and
- `get_column_plot()` for visual inspection of feature distributions.

D. Experimental Design

Two main experiments were conducted:

1) Experiment 1: Baseline Comparison with Feature Selection: The original dataset was split into training and test sets using stratified sampling to maintain class proportions. SMOTE [14] was applied to oversample the minority classes in the training set. Subsequently, both the training and test sets were standardized using *StandardScaler*.

Feature selection was performed on the training data using Recursive Feature Elimination (RFE) to identify the most informative features. The selected features were then used to train three classifiers: XGBoost [15], Logistic Regression, and Random Forest [16]. GridSearchCV was employed to perform hyperparameter tuning and select the best configuration for each model.

2) Experiment 2: Feature Selection with Synthetic Data: The original training set was augmented with synthetic data, and the combined dataset was standardized using *StandardScaler*.

Three feature selection techniques were evaluated:

- **Recursive Feature Elimination (RFE) [17]**: RFE is a wrapper method that recursively removes the least important features based on model weights or feature importances. At each step, the model is retrained, and the process continues until the desired number of features is selected.
- **Minimum Redundancy Maximum Relevance (mRMR) [18]**: The minimum redundancy is calculated using

$$\min_{S\subset\Omega} \frac{1}{|S|^2} \sum_{i,j\in S} I(f_i, f_j) \tag{1}$$

where $I(f_i, f_j)$ is the mutual information between features f_i and f_j, and $|S|$ is the number of features in set S.
The maximum relevance is measured by:

$$\max_{S\subset\Omega} \frac{1}{|S|} \sum_{i\in S} I(c, f_i) \tag{2}$$

where $c = \{c_1, c_2, \ldots, c_k\}$ are the classification variables and f_i is the i^{th} feature.
Finally, mRMR selects the optimal feature set by maximizing the following objective function:

$$\max_{S\subset\Omega} \left\{ \sum_{i\in S} I(c, f_i) - \left[\frac{1}{|S|^2} \sum_{i,j\in S} I(f_i, f_j) \right] \right\} \tag{3}$$

- **Boruta [19]**: Boruta is an all-relevant feature selection method that identifies all features useful for prediction. It creates shadow features by shuffling original features

Table I
FEATURE DESCRIPTION FOR UCI DATASET.

Feature	Count	Average ($\bar{x}$)	σ	Min. value	Q_1	Q_2	Q_3	Max. value
Age (in years)	615	47.41	10.06	19	39.00	47.00	54.00	77
Sex	615	0.61	0.49	0	0.00	1.00	1.00	1
ALB (g/L)	614	41.62	5.78	14.90	38.80	41.95	45.20	82.20
ALP (IU/L)	597	68.28	26.03	11.30	52.50	66.20	80.10	416.60
AST (U/L)	615	34.79	33.09	10.60	21.60	25.90	32.90	324.00
BIL (umol/L)	615	11.40	19.67	0.80	5.30	7.30	11.20	254.00
CHE	615	8.20	2.21	1.42	6.94	8.26	9.59	16.41
CHOL (mmol/L)	605	5.37	1.13	1.43	4.61	5.30	6.06	9.67
CREA (umol/L)	615	81.29	49.76	8.00	67.00	77.00	88.00	1079.10
CGT (IU/L)	615	39.53	54.66	4.50	15.70	23.30	40.20	650.90
PROT (g/L)	614	72.04	5.40	44.80	69.30	72.20	75.40	90.00
ALT (U/L)	614	28.45	25.47	0.90	16.40	23.00	33.08	325.30

and then compares their importance using a classifier like Random Forest. Features that consistently outperform the best shadow features are retained.

Each method was applied to the training data to identify the most informative features. Their performance was compared across the three classifiers, and a summary of the results is presented in the *Results and Discussion* section.

Among these, RFE produced the best performance and was used in subsequent evaluations. Classifiers were trained using the selected features, and model performance was compared between SMOTE-augmented and synthetic-data-augmented setups.

A visual overview of the full experimental workflow, including both experiments, is shown in Figure 1.

E. Imbalance Handling and Model Optimization

To mitigate class imbalance:

- `class_weight='balanced'` was applied to Logistic Regression and Random Forest.
- For XGBoost, the `scale_pos_weight` parameter was adjusted based on class ratios.

All models were optimized using GridSearchCV with cross-validation to ensure robust hyperparameter tuning.

IV. RESULTS

In this section, we explore and analyze the performance of different classification models under two experimental setups: one using SMOTE-augmented data and the other using synthetic data generated via SDV. We further show the effectiveness of various feature selection methods and interpret the model outputs using explainable AI techniques.

A. Finding the Best Feature Selection Method

In this experiment, we use three feature selection methods—Recursive Feature Elimination (RFE), Minimum Redundancy Maximum Relevance (MRMR), and Boruta—to optimally classify liver fibrosis staging. All three methods aim to identify the most informative features; RFE performed its task effectively.

With RFE, we successfully reduced the number of features to 6 (Age, ALP, AST, CHE, CREA, GGT), from 12, without compromising the performance of the models after SDV-based oversampling. The features selected by RFE aligned with clinically significant indicators. Liver enzymes (ALP, AST, CHE) are major indicators of liver diseases, which were also prioritized by RFE. Less predictive features like "epigastric pain" were removed. Overall, RFE performed the best over all methods and proved itself effective for balancing feature reduction with predictive performance, especially when combined with synthetic data augmentation.

Table II
COMPARISON OF FEATURE SELECTION METHODS

Method	Model	F1 (Weighted Average)	Number of Features
RFE	LR	0.97	6
MRMR	LR	0.95	8
Boruta	LR	0.94	10

B. SHAP Feature Importance Across Classes

As illustrated in Figure 2, we present the average SHAP values for each selected feature, highlighting their importance on the model's output across all fibrosis classes. Features like CHE, ALP, and AST show strong contributions to differentiating between multiple classes, particularly Class 2 and Class 3

C. Experiment 1: SMOTE-Based Training

In this experiment, we applied SMOTE to generate samples for minority classes to address the imbalance before evaluating model performance. This significantly improved F1-score and ROC-AUC across all models tested.

Among the classifiers—XGBoost, Random Forest, and Logistic Regression—Logistic Regression outperformed others with notably better scores across all minority classes. Particularly, it achieved F1-scores of 0.99, 0.80, 0.77, and 0.83 for classes 0, 1, 2, and 3, respectively, along with near-perfect ROC-AUC scores close to 1.00 for all classes.

D. Experiment 2: SDV Synthetic Data

Although SMOTE helped to decrease class imbalance by generating synthetic samples, we further explored SDV-based oversampling to enhance minority class representation. Unlike SMOTE, which synthesizes samples by averaging feature

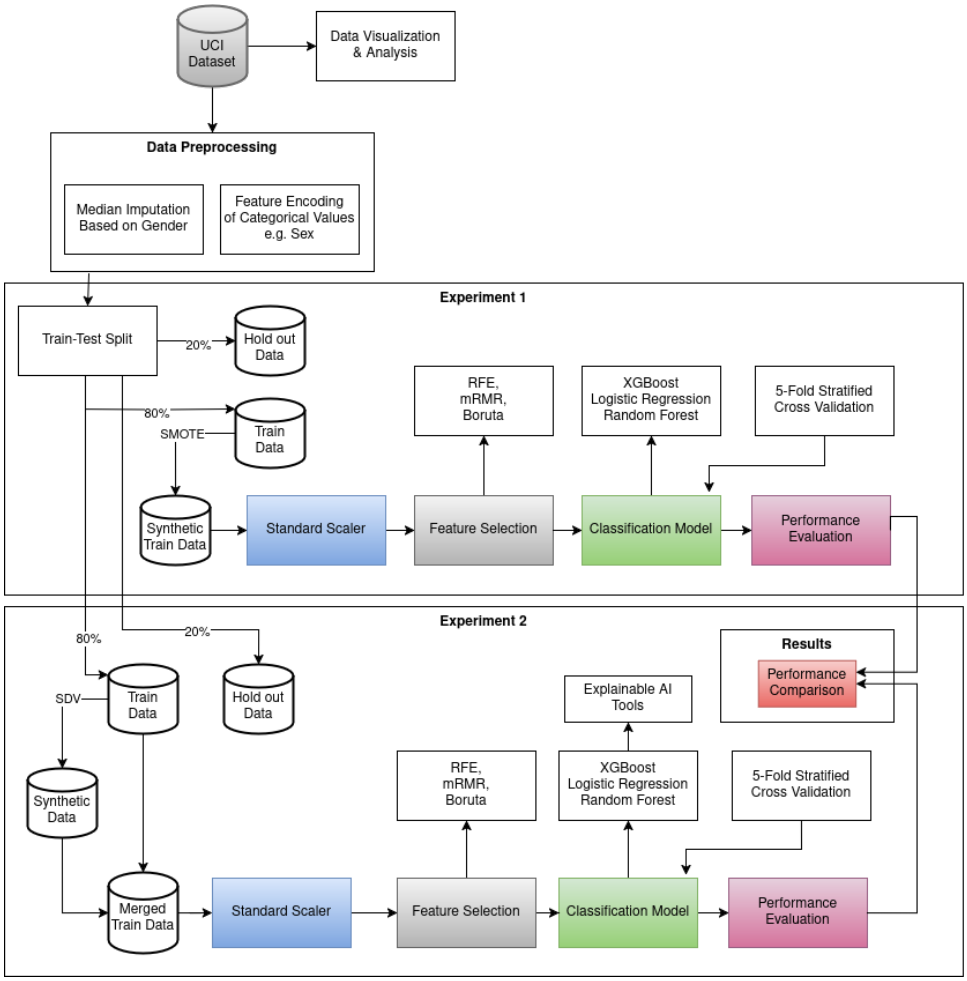

Figure 1. Overall experimental workflow for data preprocessing, synthetic data generation, feature selection, and model evaluation.

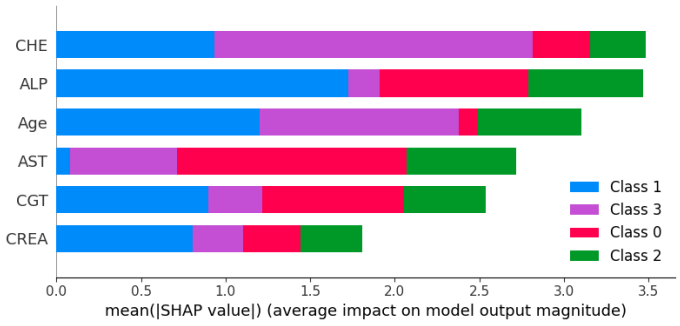

Figure 2. SHAP Feature Importance Across Classes

Table III
MODEL PERFORMANCE METRICS TRAINED ON SMOTE GENERATED SYNTHETIC DATA

Model	Class	Precision	Recall	F1	ROC
XGB	0	0.98	0.99	0.99	0.98
	1	0.67	0.40	0.50	0.95
	2	0.60	0.75	0.67	0.98
	3	0.67	0.67	0.67	0.98
RF	0	0.96	1.00	0.98	0.98
	1	1.00	0.40	0.57	0.96
	2	0.60	0.75	0.67	0.98
	3	1.00	0.67	0.80	0.98
LR	0	1.00	0.97	0.99	1.00
	1	0.80	0.80	0.80	0.99
	2	0.57	1.00	0.73	0.98
	3	0.83	0.83	0.75	0.98

space distances, SDV (Synthetic Data Vault) learns the underlying data distribution and generates more realistic, diverse samples. In this work, SDV was applied only to generate samples for the minority classes. After using this generated dataset with the training dataset, all classifiers improved performance, particularly in recall and F1-score for minority classes. Among them, Logistic Regression again achieved the best results. It confirms the effectiveness of SDV in handling complex class imbalances and maintaining decision boundary integrity.

After applying SDV-based oversampling, model performance improved significantly across all classifiers, especially for minority classes. Unlike SMOTE, SDV generated more realistic samples by modeling data distributions, which helped improve recall and F1-scores without overfitting.

Logistic Regression continued to outperform others, achiev-

Table IV
MODEL PERFORMANCE METRICS TRAINED ON SDV GENERATED SYNTHETIC DATA

Model	Class	Precision	Recall	F1	ROC
XGBoost	0	0.98	1.00	0.99	0.99
	1	0.67	0.40	0.50	0.98
	2	0.60	0.75	0.67	0.98
	3	0.83	0.71	0.77	0.99
RF	0	0.95	1.00	0.97	1.00
	1	1.00	0.20	0.33	0.99
	2	1.00	0.75	0.86	0.98
	3	0.83	0.71	0.77	0.99
LR	0	1.00	0.98	0.99	1.00
	1	0.80	0.80	0.80	0.98
	2	0.80	1.00	0.89	0.99
	3	0.75	0.86	0.80	0.98

Table V
COMPARISON OF METRICS WITHOUT AND WITH SDV

Metric	Without SDV	With SDV
Accuracy	94.2%	96.7%
Macro F1-score	0.82	0.87
Weighted F1	0.96	0.97
Class-wise F1 (Class 0)	0.99	0.99
Class-wise F1 (Class 1)	0.80	0.80
Class-wise F1 (Class 2)	0.73	0.89
Class-wise F1 (Class 3)	0.75	0.80
Macro ROC-AUC	$\sim$0.99	$\sim$0.99

ing F1-scores of 0.99, 0.80, 0.89, and 0.80 for classes 0 to 3, respectively. It also maintained high ROC-AUC values (≥ 0.98). Random Forest and XGBoost also showed gains, particularly for class 2 and class 3. The average F1-score increased to 0.86, and ROC-AUC to 0.98, confirming SDV's effectiveness in addressing class imbalance.

The application of SDV-based oversampling led to notable improvements in the performance of the Logistic Regression model, particularly in handling minority classes. The most significant gain was observed in Class 2, where the F1-score increased from 0.73 to 0.89, indicating enhanced precision and recall for this underrepresented group. While both models maintained near-perfect ROC-AUC values, the SDV-enhanced model demonstrated more consistent performance across all classes, especially for the minority categories. Furthermore, macro-averaged F1-score improved from 0.84 to 0.87, and overall accuracy rose from 95.9% to 96.7%, highlighting the effectiveness of SDV in achieving better class balance and overall model generalization.

Logistic Regression with just 6 features performed best after SDV oversampling. The confusion matrix shows strong class-wise accuracy, especially for minority classes.

V. CONCLUSION & FUTURE WORK

A. Conclusion

In this study, we addressed the challenge of multiclass classification on highly imbalanced liver disease data by generating synthetic data using SDV combined with recursive feature elimination. This approach led to significant improvements in classification performance, particularly for minority classes that are often underrepresented in imbalanced datasets. The Logistic Regression classifier trained on the SDV-augmented dataset achieved an impressive overall accuracy of 96.73%, a macro-averaged F1-score of 0.87, and a weighted F1-score of 0.97, making it the best-performing model. Notably, it attained a near-perfect ROC-AUC score of 1.00, indicating excellent discrimination capability across all classes—performance levels rarely achieved with other methods. These results demonstrate the effectiveness of combining feature selection with high-quality synthetic data generation in addressing data imbalance and enhancing model generalization. Although the dataset used is not recent, it remains a well-established benchmark, and our proposed approach can be readily applied to modern clinical data.

B. Future Work

Building on the promising results of this study, future work will explore several avenues to enhance the robustness and generalizability of liver fibrosis classification: Development of a New Dataset: The Data set we worked is not recent. So we can collect data from multiple patients and create a new dataset. If we can create a new dataset then it will focusing on clinically relevant features identified in this study. This dataset will aim to capture more diverse patient profiles. Hybrid Dataset Construction: This dataset have comparatively less data sample. That result the dataset highly imbalanced. To address this problem, a hybrid dataset approach will be adopted by combining the newly created dataset with other publicly available liver-related medical datasets. Deep Learning Implementation: So far we apply ML model to this dataset. With a larger and more diverse dataset, we plan to experiment with deep learning models,. These models are well known for their ability to learn complex, non-linear patterns in liver fibrosis progression that may not be fully captured by traditional classifiers.

REFERENCES

[1] World Health Organization, "Hepatitis C," Apr. 2024. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c. Accessed May 10, 2025.

[2] M. Al-Mahtab, S. Rahman, F. Karim, G. Foster, and S. Solaiman, "Epidemiology of Hepatitis C Virus in Bangladeshi General Population," *Bangabandhu Sheikh Mujib Medical University Journal*, vol. 2, pp. 14–17, Nov. 2009.

[3] D. Lavanchy, "Evolving epidemiology of hepatitis C virus," *Clinical Microbiology and Infection*, vol. 17, pp. 107–115, Feb. 2011.

[4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, pp. 20–29, June 2004.

[5] H. Mamdouh, M. Y. Shams, and T. A. El-Hafeez, "Hepatitis C Virus Prediction Based on Machine Learning Framework: a Real-world Case Study in Egypt," Jan. 2022.

[6] A. Sharma, T. Khade, and S. M. Satapathy, "A cross dataset meta-model for hepatitis C detection using multi-dimensional pre-clustering," *Scientific Reports*, vol. 15, p. 7278, Mar. 2025.

[7] A. Alizargar, Y.-L. Chang, and T.-H. Tan, "Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques," *Bioengineering*, vol. 10, p. 481, Apr. 2023.

[8] T.-H. S. Li, H.-J. Chiu, and P.-H. Kuo, "Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm," *IEEE Access*, vol. 10, pp. 91045–91058, 2022.

[9] L. Chen, P. Ji, and Y. Ma, "Machine Learning Model for Hepatitis C Diagnosis Customized to Each Patient," *IEEE Access*, vol. 10, pp. 106655–106672, 2022.
[10] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—a case study," *Journal of Laboratory and Precision Medicine*, vol. 3, pp. 58–58, June 2018.
[11] K. A. Gebo, F. H. Herlong, M. S. Torbenson, M. W. Jenckes, G. Chander, K. G. Ghanem, S. S. El-Kamary, M. Sulkowski, and E. B. Bass, "Role of liver biopsy in management of chronic hepatitis C: A systematic review," *Hepatology*, vol. 36, pp. s161–s172, Nov. 2002.
[12] K. F. Lichtinghagen, Ralf and G. Hoffmann, "HCV data." UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5D612.
[13] N. N. R. . Patki, *The Synthetic Data Vault : generative modeling for relational databases*. Thesis, Massachusetts Institute of Technology, 2016. Accepted: 2017-06-06T18:44:28Z Journal Abbreviation: SDV : generative modeling for relational databases.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.
[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (San Francisco California USA), pp. 785–794, ACM, Aug. 2016.
[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.
[18] H. Peng, C. Ding, and F. Long, "Minimum Redundancy– Maximum Relevance Feature Selection,"
[19] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta – A System for Feature Selection," *Fundamenta Informaticae*, vol. 101, pp. 271–285, July 2010.