HEPATITIS C VIRUS PREDICTION IN MACHINE LEARNING

ALHASAN SALIH IBRAHIM IBRAHIM

JULY 2025

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL
COMPUTER ENGINEERING DEPARTMENT
INFORMATION TECHNOLOGY MASTER'S THESIS

HEPATITIS C VIRUS PREDICTION IN MACHINE LEARNING

ALHASAN SALIH IBRAHIM IBRAHIM

JULY 2025

ABSTRACT

HEPATITIS C VIRUS PREDICTION IN MACHINE LEARNING

IBRAHIM, ALHASAN SALIH IBRAHIM

INFORMATION TECHNOLOGY MASTER'S THESIS

Supervisor: Assist. Prof. Dr. ABDÜL KADIR GÖRÜR

July 2025, 74 Pages

The infection of the hepatitis C virus is a considerable medical field challenge globally that can require the development of effective as well as accurate diagnostic approaches. Traditional diagnostic techniques, while widely used, often have limits when it comes to accuracy, accessibility, and cost-effectiveness. This study proposes a predictive model utilizing machine learning to early diagnose the liver HCV, utilizing the Extra Trees Classifier in conjunction with the Synthetic Minority Over-Sampling Technique to address the challenge of class imbalance within the dataset. Three freely accessible datasets, HCV-EGY, ILPD, and HCV, have been used in both training and evaluation, thereby ensuring robustness and generalisability across diverse population groups. The model of this study achieves an accuracy of 98% of both the HCV and HCV-EGY datasets, while the ILPD achieved 95%. exceeding the performance of traditional diagnostic methods and demonstrating the effectiveness of machine learning in improving early HCV detection. An analysis of feature importance was performed to determine the key biomarkers that significantly influence the classification process. The interpretability component is essential, offering insights into the biological markers linked to HCV infection, which may assist in refining diagnostic criteria and treatment strategies. This study highlights the potential of non-invasive, data-driven diagnostic methods in clinical settings through the application of advanced machine learning techniques. The results indicate that machine learning models can function as dependable, efficient, and interpretable instruments to aid healthcare professionals in the early diagnosis of HCV. This research enhances theexisting evidence for AI-driven methodologies in medical diagnostics, facilitating the development of more accurate and accessible disease detection frameworks.

Keywords: Hepatitis C Virus; Liver Disease; Machine Learning; Feature Selection; Classification; SMOTE.

ÖZET

HEPATİTİS C VİRÜSÜ (HCV) MAKİNE ÖĞRENİMİ TEKNİKLERİ KULLANARAK TAHMİNİ

IBRAHIM, ALHASAN SALIH IBRAHIM
BİLGİ TEKNOLOJİLERİ YÜKSEK LİSANS TEZİ

Danışman: Dr. Öğr. Üyesi ABDÜL KADIR GÖRÜR
Temmuz 2025, 74 Sayfa

Hepatit C virüsünün enfeksiyonu, etkili ve doğru tanı yaklaşımlarının geliştirilmesini gerektirebilecek küresel ölçekte önemli bir tıbbi alan zorluğudur. Geleneksel tanı teknikleri, yaygın olarak kullanılsa da, genellikle doğruluk, erişilebilirlik ve maliyet etkinliği açısından sınırlamalara sahiptir. Bu çalışma, karaciğer HCV'sinin erken teşhisi için makine öğrenimini kullanan bir tahmin modeli önermektedir; veri kümesindeki sınıf dengesizliği sorununu ele almak için Ekstra Ağaçlar Sınıflandırıcısı ile Sentetik Azınlık Aşırı Örnekleme Tekniği bir arada kullanılmaktadır. Üç serbest erişilebilir veri seti, HCV-EGY, ILPD, HCV, hem eğitim hem de değerlendirme için kullanılmıştır, böylece çeşitli nüfus grupları arasında sağlamlık ve genelleştirilebilirlik sağlanmıştır. Bu çalışmanın modeli, hem HCV hem de HCV-EGY veri setlerinde %98 doğruluk elde ederken, ILPD %95 doğruluk elde etmiştir. geleneksel tanı yöntemlerinin performansını aşarak, erken HCV tespitini iyileştirmede makine öğreniminin etkinliğini göstermektedir. Özellik önemliliği analizi, sınıflandırma sürecini önemli ölçüde etkileyen ana biyomarkerleri belirlemek için gerçekleştirildi. Yorumlanabilirlik bileşeni, HCV enfeksiyonu ile bağlantılı biyolojik belirteçler hakkında içgörüler sunarak, tanı kriterlerinin ve tedavi stratejilerinin geliştirilmesine yardımcı olabilir. Bu çalışma, ileri düzey makine öğrenimi tekniklerinin uygulanması yoluyla klinik ortamlarda invaziv olmayan, veri odaklı tanı yöntemlerinin potansiyelini vurgulamaktadır. Sonuçlar, makine öğrenimi modellerinin HCV'nin erken teşhisinde sağlık profesyonellerine yardımcı olmak için güvenilir, verimli ve yorumlanabilir araçlar olarak işlev görebileceğini göstermektedir. Bu araştırma, tıbbi teşhislerde yapay zeka destekli metodolojiler için mevcut kanıtları güçlendirerek, daha doğru ve erişilebilir hastalık tespit çerçevelerinin geliştirilmesini kolaylaştırmaktadır.

Anahtar Kelimeler: Hepatit C virüsü; karaciğer hastalığı; Makine öğrenimi; Özellik seçimi; Sınıflandırma; SMOTE.

ACKNOWLEDGEMENT

The only thing outweighing my exhaustion from this journey is the deep gratitude I feel for every single person who stood by my side during the making of this thesis.

My dear supervisor, **Dr**.**Abdül Kadir Görür**. There are so many wonderful memories and inspiring conversations that I will treasure forever. Your support and belief in me were never a surprise, and there’s nothing I can say that hasn’t already been said about your incredible kindness. But what touched me most was your humility. You have been a constant reminder of everything that is beautiful about collaboration, learning. Your mentorship, expertise and encouragement have been instrumental in shaping not only this work but also the person I am becoming.

To my father, **Mr. Salih Ibrahim**, and my mother, **Mrs. Hanan Khalaf**, from the bottom of my heart: thank you for every sacrifice, every prayer, every piece of advice. You are the pillars of my life, and this achievement belongs as much to you as it does to me.

Special friends, **Ali Asal** and **Anas Jalal**. Thank you for every laugh, every late-night chat, and every shared moment that made this journey lighter and more meaningful. Your friendship has been a source of joy throughout this entire journey. Thank you truly for walking this path with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

DAA	: Direct Acting Antivirals
AI	: Artificial Intelligence
PCA	: Principal Component Analysis
HCC	: Hepatocellular Carcinoma
HCV	: Hepatitis C Virus
ETC	: Extra Tree Classifier
RNA	: Ribonucleic Acid
KNN	: K-nearest Neighbor
RF	: Random Forest
DT	: Decision Tree
ADT	: Alternative Decision Tree
SMOTE	: Synthetic Minority Over-sampling Technique
PF	: Polynomial Feature
AD	: Average Derivative
LR	: Logistic Regression
NB	: Naïve Bayes
SFS	: Sequential Forword
SVM	: Support Vector Machines
UCI	: University of California, Irvine
ANN	: Artificial Neural Network
CHC	: Chronic Hepatitis C
MSE	: Mean Square Error
ML	: Machine Learning
ILPD	: Indian Liver Patent Dataset
GBM	: Gradient Boosting Machine

CHAPTER I
INTRODUCTION

1.1 OVERVIEW

The global health concern of Hepatitis C Virus (HCV) is substantial, with an estimated prevalence in approximately 3% of the global population and an estimated 170 million individuals affected worldwide [1]. As a blood-borne pathogen, HCV can cause persistent liver diseases such the cirrhosis and hepatocellular cancer, and also the failure of liver if left untreated [2]. Despite advancements in diagnostic methods, early detection remains a difficulty due to the asymptomatic nature of infection during its initial stages [3]. The progression from acute to chronic infection often goes unnoticed, complicating timely intervention. Current diagnostic tools, while effective, face limitations in sensitivity and specificity, especially in resource-limited settings [4].

In recent years, machine learning has emerged as an innovative technology within the field of artificial intelligence, with applications that extend diverse domains, including healthcare [5]. Machine learning algorithms, especially supervised learning methods like the decision trees and support vector machine, neural networks are effective in recognizing intricate patterns in extensive datasets [6]. These algorithms analyze multiple variables concurrently, reveal concealed relationships, and produce predictive models with significant accuracy [7]. The medical field has to enable the analysis of healthcare data, machine learning can be used, as it is a tool capable of handling large amounts of data.

The integration of the machine learning with healthcare demonstrated significant outcomes in the fields of the disease predictions and diagnosis [8]. By using the clinical and demographic, and laboratory data the machine learning models could pretty improve the early detection of the diseases like HCV, that could pretty help the clinicians in making informed by data-driven decisions [9]. This predictive capability is the most essential because it could enhance the patient outcomes and also for streamlining the allocation of the healthcare resources. the healthcare systems are gradually adopting data-driven approaches, machine learning emerges as a partner in addressing the public health challenges, offering the potential for personalized medicine and more efficient disease management [10].

1.2 HEPATITIS C VIRUS

One of the most massive solid organs in the human body is the liver , the biggest gland, and one of the most important organs [11]. The Hepatitis C Virus (HCV) is a blood-borne, single-stranded RNA virus belonging to the Flaviviridae family, mostly impacting the liver, causing both acute and chronic infections [12]. Less frequently, HIV can be passed from mother to child or through sexual contact, but the most common way it spreads is by exposure to infected blood, which can happen through injection drug use or improper medical practices [13]. Chronic hepatitis C virus infection can lead to a variety of serious liver problems, including cirrhosis and hepatocellular cancer [14]. The process of diagnosis entails the identification of anti-HCV antibodies and the confirmation of active infection through HCV RNA assays [15]. The advent of direct-acting antivirals (DAAs) has markedly enhanced treatment outcomes, achieving cure rates exceeding 95% with reduced side effects, global access to diagnosis and treatment continues to pose challenges.

1.3 MACHINE LEARNING

Machine learning is a subset of artificial intelligence (AI) that focusses on creating algorithms and models enabling computers to learn from data and make predictions or decisions, without explicit programming. ML's main goal is to find patterns in data and use them to guess what will happen in the future, automate jobs, or help people make decisions [16]. ML techniques are broadly classified into three types: supervised learning, unsupervised learning and reinforcement learning, where agents learn optimal actions through trial and error while receiving feedback from their environment [17][18]. Decision trees, support vector machines (SVM), and neural networks are examples of algorithms that are frequently utilized in supervised learning processes for the purpose of performing classification and regression problems [19]. Unsupervised learning, on the other hand, makes use of k-means in order to discover data structures that do not have any established labels. Reinforcement learning draws from behavioral psychology and is utilized in dynamic settings, such as robots and game creation.

Healthcare, banking, marketing, and autonomous systems are just a few of the many sectors that can benefit from ML's revolutionary applications. Disease diagnosis, patient risk assessment, and drug discovery are just a few areas where ML models have proven useful in healthcare. Despite its potential, machine learning has obstacles such as data privacy concerns, interpretability of complicated models, and the risk of algorithmic bias.

1.4 MACHINE LEARNING IN HCV PREDICTION

The medical field has rapidly integrated machine learning, significantly impacting diagnosis, treatment, ML methods can examine complex collections of information and identify hidden patterns which might ignore traditional statistical methods. In HCV field, ML methods are most commonly used in early diagnosis, disease classification, treatment response prediction, and risk stratification of patients [20].

There's an important place of machine learning models in prediction of the progression of HCV related liver diseases such as cirrhotic patients and those with hepatocellular carcinoma (HCC) [7]. These predictive abilities allow doctors to intervene promptly, which reduces the risk of serious complications. ML also makes personalized treatment plans possible by predicting how every single patient will respond to antiviral medications, which increases the effectiveness of the treatment [20].

In addition to its therapeutic uses, machine learning also contributes to epidemiological surveillance, enabling public health officials to track the transmission trends of the disease and assess the effectiveness of intervention plans. In addition to increasing diagnostic accuracy, The use of machine learning in healthcare can lower healthcare expenses, improve operational efficiency, and encourage decision-making based on evidence [20].

1.5 PROBLEM STATEMENT

Hepatitis C Virus (HCV) is a serious global health problem. It infects millions of people around the world each year. HCV can result in severe liver consequences if it remains unnoticed and untreated. Traditional HCV diagnostic procedures are slow and costly and may not detect infection during its early stages. What's more, the fact that HCV is asymptomatic in its early stages means that the disease gets diagnosed with a delay. This raises risk of progression and transmission of the disease. The limitation of the above models calls for new data-driven tools that could help improve early identification, risk assessment, and treatment optimization.

By systematically analyzing extensive of clinical and biochemical data in search of patterns that more conventional diagnostic methods might miss, machine learning (ML) shows great promise in solving these problems. Model accuracy, generalizability, and interpretability are three obstacles to ML's effective application in HCV diagnosis and prognosis. Also, to make sure ML models are reliable and well-received by healthcare providers, they need to be rigorously validated before being integrated into clinical workflows.

1.6 RESEARCH AIM

This study will aid in the advancement of AI-assisted healthcare tools, promoting more effective disease management and treatment planning for HCV patients. Research objectives include:

- Develop robust machine learning models for predicting HCV progression, focusing on improving diagnostic accuracy and generalizability across diverse patient populations.
- Address class imbalance issues inherent in medical datasets by applying advanced resampling techniques ensuring fair and balanced learning across all disease categories.
- Identify and rank critical prognostic markers (clinical and biochemical features) using ensemble-based algorithms, enabling interpretability and supporting clinicians in early detection and treatment planning.
- Evaluate and compare the performance of ensemble learning classifiers on balanced dataset, quantifying the impact of class balancing on model reliability and predictive power.

1.7 RESEARCH CONTRIBUTIONS

This study makes several significant contributions to the field of AI-assisted healthcare. It develops a comprehensive machine learning pipeline that incorporates advanced data preprocessing, feature selection, and class balancing techniques specifically tailored for HCV datasets. By identifying critical prognostic markers from clinical and biochemical data, the work provides actionable insights to support physicians in making informed decisions.

The proposed framework is scalable and interpretable, making it adaptable for diagnosing other liver-related diseases and broadening its application in medical diagnostics. Furthermore, it promotes early diagnosis and intervention strategies, which are crucial for improving treatment planning and enhancing patient outcomes. This integration of machine learning techniques strengthens the potential for AI-driven tools in healthcare, addressing key challenges like data imbalance and feature redundancy while ensuring robust and reliable predictions. Overall, the study bridges technological advancements with clinical needs to support better disease management.

1.8 RESEARCH QUESTION

This thesis is organized around the following guiding research questions:

1. How efficient is the performance of clinical and biological feature-based ML models in predicting Hepatitis C Virus infection?
2. How does the consideration of class imbalance in HCV prediction models impact their accuracy and medical applicability?
3. How does the choice of pertinent biomarkers features impact the explain ability and the diagnostic accuracy of the proposed machine learning models within the context of their evaluation?
4. Is it feasible to create a machine learning model on HCV-related datasets of differing characteristics and population diversity that would maintain a strong performance across the diverse datasets?

1.9 STRUCTURE OF THESIS

The research anticipates the proposed advancements in diagnostic methodologies and data analysis will contribute to improved patient outcomes and optimized resource allocation in medical practice. Upcoming chapters include:

- Chapter II (Literature Review) provides a comprehensive overview of recent studies focused on HCV prediction, highlighting various machine learning approaches, datasets, and challenges such as class imbalance and feature selection. It critically examines the strengths and limitations of existing models to identify research gaps.

- Chapter III (Methodology) details of proposed methodology in this research, emphasizing the Extra Trees Classifier (ETC) model and its advantages for HCV prediction. It describes data preprocessing steps, including handling class imbalance and feature selection strategies, ensuring robustness and interpretability. The chapter also outlines the evaluation metrics and validation protocols used to assess model performance.
- Chapter IV (Experiments and Results) presents and analyzes the results in all datasets, comparing the performance of ETC against baseline models, and discusses the impact of used techniques on predictive accuracy.
- Chapter V (Conclusion and Future Work) summarizes the study's key contributions and findings. It also focuses on the clinical significance and practical application of the proposed model, while clarifying the study's limitations and potential for future development.

CHAPTER II
LITERATURE REVIEW

2.1 INTRODUCTION

HCV remains a major global health issue due to liver complications and asymptomatic early stages. WHO estimates 71 million are chronically infected, with regional prevalence disparities [65]. In Turkey, anti-HCV seroprevalence is 0.95-2.4%; cirrhosis and hepatocellular carcinoma are severe outcomes [63]. A 2023 Iraqi study found a 3.2% HCV prevalence overall, with higher rates in thalassemia (15.45%) and renal failure (6.66%) patients, highlighting the need for better screening [64]. Conventional methods for diagnosing liver fibrosis stages, such as liver biopsy, are invasive and costly, prompting researchers to explore non-invasive techniques. Machine learning has emerged as a promising tool in this regard, offering high accuracy in detecting liver fibrosis stages and predicting HCV progression. Several studies have leveraged ML algorithms to improve the classification of HCV stages based on patient data. Early works largely employed decision tree-based models, but recent efforts have expanded the use of various classifiers, including K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM), often coupled with feature selection techniques to enhance performance. These studies aim to develop reliable medical decision support systems for diagnosing HCV.

2.2 REVOLUTIONIZING NON-INVASIVE TREATMENT STRATEGIES FOR LIVER DISEASES THROUGH MACHINE LEARNING

Viral hepatitis can contribute to morbidity and mortality in patients who are receiving chemotherapy [21]. Machine learning (ML) significantly enhances non-invasive treatment approaches by providing advanced analytical tools that accurately assess liver diseases, mitigating the need for invasive procedures. As illustrated in the Figure 1. A biopsy is typically advised for the initial evaluation of patients with persistent HCV infection. It is beneficial for assessing disease severity (fibrosis stage) and for evaluating the extent of necrosis and inflammation [22]

AI models treat predictions as potential biomarkers, synthesizing a diverse array of clinical, omics, and imaging data to predict disease outcomes and severity. while emphasizing the complexity of interactions that can occur at the molecular level.

By leveraging large datasets, ML models can discern intricate patterns within high-dimensional data, enabling the early detection of liver conditions such as hepatic steatosis, hepatocellular carcinoma (HCC), ML algorithms can identify subtle abnormalities indicative of disease progression without the discomfort and risks associated with invasive methods like biopsies, The emphasis on non-invasive strategies not only enhances patient comfort and safety but also significantly reduces healthcare costs associated with surgical interventions and hospital stays. By streamlining the diagnosis and treatment process, ML enforces a more efficient healthcare system, reaffirming its critical role in advancing non-invasive treatment strategies in liver disease management [23].

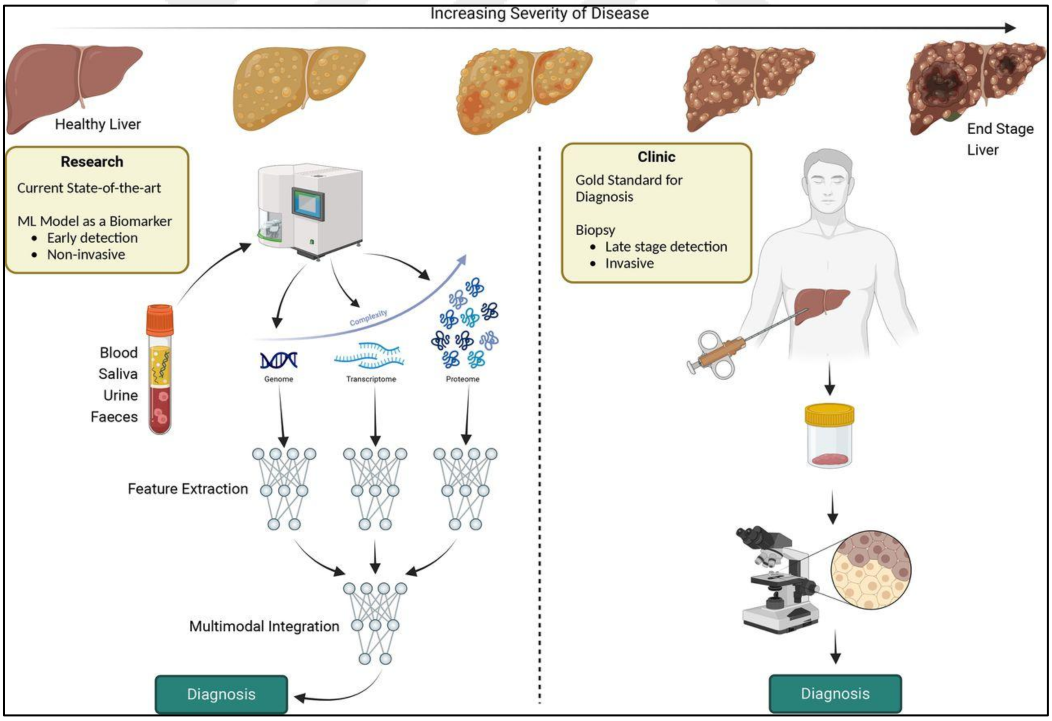


Figure 1: ML in LD predicting [23]

2.3 BIOMARKERS AND DATASETS IN HEPATOCELLULAR CARCINOMA RESEARCH

Biomarkers serve as a crucial source of datasets in medical research; biomarker data can play an integral role in AI as well as machine learning applications. As shown in Figure 2, the correlation among artificial intelligence, machine learning, and deep learning, these biomarkers are instrumental in training predictive models and validating their efficacy. Genomic data, for instance, can refine machine learning models to improve prognostic predictions, whereas imaging biomarkers may increase the precision of lesion characterization in radiological diagnostic*s* [24].

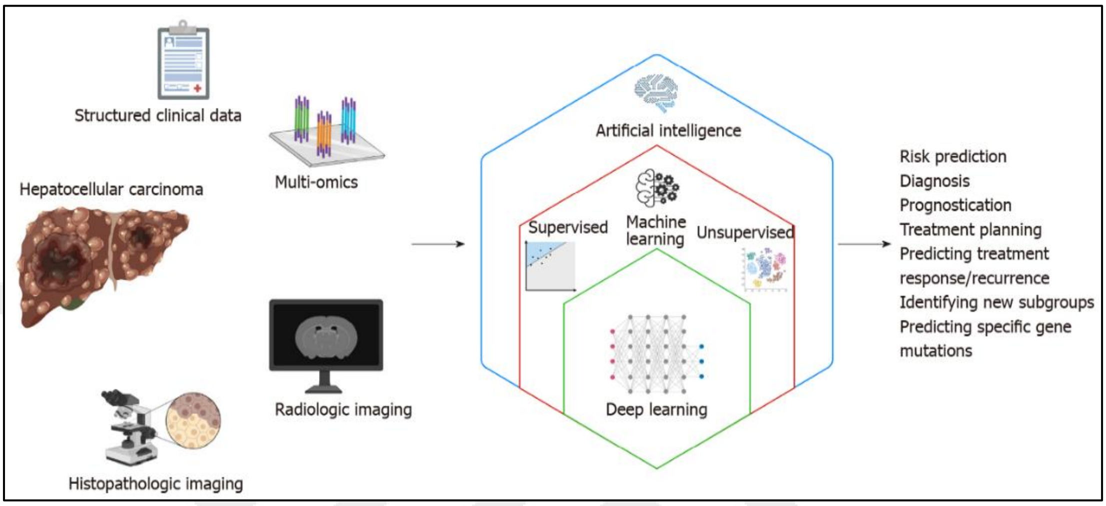


Figure 2: Deferent sources of datasets [24]

2.4 RELATED WORK

Hepatitis C virus (HCV) poses a significant global health challenge, affecting an estimated 120–130 million people worldwide, or approximately 3% of the population [65]. Without timely treatment, acute HCV infections often progress to chronic hepatitis C virus (CHC) conditions, leading to severe liver diseases such as cirrhosis and hepatocellular carcinoma. In recent years, machine learning (ML) algorithms have emerged as effective tools for predicting and classifying stages of liver fibrosis in HCV patients due to their ability to model the non-linear progression of the disease. However, accuracy levels reported across studies vary widely due to differences in datasets, feature engineering strategies, class balancing approaches, and classification techniques. Advanced ML methods combined with robust feature selection mechanisms and explainable artificial intelligence (XAI) have demonstrated notable potential in improving diagnostic precision while addressing challenges such as class imbalance, data heterogeneity, and interpretability.

Konerman et al. (2019) [25] pioneered the use of longitudinal boosted-survival-tree models to predict cirrhosis development in CHC patients using data from the Veterans Health Administration (VHA) Corporate Data Warehouse. The study analyzed 72,683 patients between 2000 and 2016, excluding individuals with pre-existing cirrhosis or initial aspartate aminotransferase-to-platelet ratio index (APRI) > 2. Longitudinal models incorporating dynamic variables (e.g., slopes and variability) outperformed cross-sectional (CS) models, achieving a concordance index of 0.774 and superior area under the receiver operating characteristic curve (AUC-ROC) scores at 1-, 3-, and 5-years post-baseline. Similarly, Barakat et al. (2019) [26] enhanced pediatric CHC fibrosis prediction using Random Forest (RF) classifiers in combination with APRI and FIB-4 (fibrosis-4 index) tailored for children. Their dataset of 166 Egyptian children achieved AUCs of 0.903, 0.894, and 0.822 for various fibrosis categories, highlighting the strength of RF in non-invasive diagnostics despite challenges such as data imbalance in advanced fibrosis stages.

In 2020, Ahammed et al. [27] proposed a classification model for fibrosis staging in 1,385 Egyptian patients (HCV-EGY dataset) using K-nearest neighbors (KNN), which achieved 94.40% accuracy after applying the Synthetic Minority Oversampling Technique (SMOTE) to mitigate class imbalance. Nandipati et al. (2020) [28] further investigated multi-class (F1–F4) and binary classification on the same dataset using KNN, RF, Support Vector Machine (SVM), neural networks (NN), and Naïve Bayes (NB), noting that RF achieved higher accuracy (54.56%) for binary labels, while KNN performed better for multi-class labels.

By 2021, studies expanded into novel methodologies and datasets. Syafaah et al. [29] evaluated NN, RF, NB, and KNN classifiers on a small cohort of 73 patients, reporting the highest accuracy (95.12%) with NN models, while Ghazal [30] introduced the Hep-Pred system using fine Gaussian SVM on the HCV-EGY dataset, achieving 97.9% accuracy through five-fold cross-validation. Mostafa et al. (2021) [31] focused on dimensionality reduction via principal component analysis (PCA) and handling missing data with multiple imputations by chained equations (MICE), where RF achieved 98.14% accuracy on a 615-patient dataset. Butt et al. (2021) [32] proposed the Intelligent Hepatitis C Stage Diagnosis System (IHSDS), powered by artificial neural networks (ANN), reaching validation precision of 94.44%. Peng et al. (2021) [33] adopted an XAI framework combining SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Partial Dependence Plots (PDP), achieving 91.9% accuracy with RF on a benchmark hepatitis dataset, effectively balancing model accuracy and interpretability.

Kaunang (2022) [34] tested six ML algorithms on 589 patient records, where logistic regression (LR) achieved the highest accuracy (97.9%) despite limitations due to class imbalance. Shinde et al. (2022) [35] developed a Clinical Decision Support System (CDSS) using the Indian Liver Patient Dataset (ILPD), where RF models combined with oversampling and grid search demonstrated high predictive efficiency. Straw and Wu (2022) [36] examined gender biases in healthcare algorithms using ILPD, revealing disparities in false negative rates between male and female patients, thus emphasizing the need for demographic equity assessments.

Recent advancements in 2023 focused on ensemble learning and interpretability. Sachdeva et al. [37] applied SMOTE to improve classifier performance on the UCI hepatitis dataset, with LR achieving 93.18% accuracy. Kim et al. [38] targeted HCV detection in diabetic patients using NHANES data, where Least Absolute Shrinkage and Selection Operator (LASSO) models achieved an AUC-ROC of 0.810. Alotaibi et al. [39] used ensemble models including Extra Trees, which achieved 96.92% accuracy, while Ali et al. [40] applied Sequential Forward Selection (SFS) and SHAP on Jordan University Hospital data to interpret diagnostic predictions effectively. Amin et al. [41] combined PCA, Factor Analysis (FA), and Linear Discriminant Analysis (LDA) within an ML framework, achieving 88.10% accuracy on ILPD. Md et al. [42] introduced advanced preprocessing (e.g., multivariate inference, log transformation) and used ensemble models such as XGBoost and Extra Trees, achieving 91.82% accuracy.

In 2024, Karna et al. [43] explored dimensionality reduction using both linear techniques (LDA, FA) and non-linear techniques (t-SNE, UMAP) to improve fibrosis staging, achieving 98.31% accuracy in 10-fold cross-validation with RF. Finally, Dashti et al. [44] introduced a fully automated diagnostic model using multilayer perceptron (MLP) neural networks, achieving a remarkable 99.5% accuracy on ILPD, demonstrating the potential of reverse learning techniques in early liver disease detection.

Recent studies, as shown in Table1, on HCV and liver disease prediction have employed diverse machine learning techniques, datasets, and feature engineering strategies, achieving accuracies ranging from 54.56% to 99.5%. While neural networks and SVMs demonstrated high performance, limitations such as small datasets, class imbalance, and interpretability issues highlight the need for robust, generalizable, and explainable models.

Table 1: Comparing related work

Ref	Dataset Used	Aim	Methods	Results	Weaknesses
[26]	HCV-EGY Data	Pediatric HCV fibrosis staging	RF with APRI and FIB-4 cutoffs	AUC: 0.903	Imbalance in advanced fibrosis cases
[27]	HCV-EGY Data	Fibrosis staging	SMOTE + KNN	94.40%	Few features used
[28]	HCV-EGY Data	HCV stage prediction	Feature selection + ML classifiers	54.56% (RF)	Low multi-class prediction accuracy
[29]	HCV Data	HCV detection	KNN, NB, NN, RF	95.12% (NN)	Neural networks require intensive tuning
[30]	HCV-EGY Data	HCV staging	Fine Gaussian SVM	97.9%	SVM is computationally expensive and less interpretable
[31]	HCV-EGY Data	Liver disease predictor extraction	PCA, SMOTE + RF	98.14%	Limited features and data size
[32]	HCV-EGY Data	HCV staging	ANN	98.89%	Potential overfitting on validation
[33]	Hepatitis Data	Explainable AI for hepatitis diagnosis	SHAP, LIME + RF	91.9%	Relatively small dataset
[34]	ILPD	HCV classification	Logistic Regression, others	97.9% (LR)	Class imbalance issues
[35]	ILPD	Early liver disease diagnosis	DT, RF, NB, SVM	99.5% (RF)	Small dataset size limited use of ensemble models
[36]	ILPD	Gender bias analysis in liver prediction	Sex-stratified ML (LR, RF, SVM, NB)	79.40% (SVM)	Focused more on fairness than optimization
[37]	Hepatitis Data	Systematic diagnostic strategy	SMOTE + LR, RF, SVM, KNN	93.18% (LR)	Limited number of records
[38]	NHANES Data	Hepatitis in diabetic patients	LASSO, RF, SVM, XGBoost	AUC: 0.810	Imbalanced data

Table 1 Cont.

Ref	Dataset Used	Aim	Methods	Results	Weaknesses
[39]	HCV-EGY Data	Cirrhosis detection	Extra Trees + SHAP, LIME	96.92%	Interpretability still relies on external tools
[40]	Jordan Hospital	Chronic liver disease diagnosis	SFS + SMOTE + ML classifiers	83% avg	Small dataset
[41]	ILPD	Chronic liver disease classification	PCA + FA + LDA + ML classifiers	88.10%	Did not address class imbalance effectively
[42]	ILPD	Advanced preprocessing for liver disease	Scaling + ETC, RF, XGBoost	91.82% (ETC)	Slight overfitting risk
[43]	ILPD	Dimensionality reduction + ML	t-SNE, UMAP + RF, KNN, MLP, LR	98.31% (10-fold)	t-SNE, UMAP not ideal for predictive modeling
[44]	ILPD	Self-predictive diagnostic system	MLP Neural Network	99.5%	Potential overfitting

2.5 DATASETS USED IN STUDIES

Studies utilized diverse datasets varying in size, feature types, and population coverage, as shown in Table 2. The HCV-EGY dataset was the most widely used, focusing on Egyptian patients and offering rich clinical and biochemical features, but limited in geographic diversity. The ILPD dataset, though smaller, was frequently applied for liver disease prediction and highlighted issues of missing values and class imbalance.

Table 2: Comparison of data sets used in previous studies

Dataset	Studies	Size	Feature Types	Limitations
HCV-EGY Data	[27], [28], [30], [32], [39]	1,385 patients	Demographic, clinical, biochemical	- Egyptian population - Potential class imbalance
HCV Data	[29]	73 patients	Biochemical, serological, histopathology	- Small size - Pediatric population focus

Table 2 cont.

Dataset	Studies	Size	Feature Types	Limitations
Hepatitis Data	[33], [37]	155 patients	Demographic, biochemical	- Small dataset - Imbalanced classes
ILPD	[34], [35], [36], [41], [42], [43], [44]	583 patients	Demographic, clinical, biochemical	- Missing values Class imbalance - Regional bias (India)
NHANES Data	[38]	~10,000 records (subset used)	Demographic, anthropometric, clinical	- Not specific to hepatitis - Heterogeneo us data sources
Jordan Hospital	[40]	1,801 patients	Biochemical, demographic	- Regional specificity (Jordan)

2.6 RESEARCH GAPS

Despite significant advancements in machine learning applications for HCV prediction and staging, several critical research gaps remain unaddressed. Many existing studies relied on imbalanced datasets where advanced stages of fibrosis or cirrhosis were underrepresented, leading to biased models that perform poorly on minority classes. This imbalance compromises diagnostic accuracy for high-risk patients, which is clinically unacceptable. Additionally, while techniques like SMOTE and class weighting have been used, they were often applied superficially without integrating them into a comprehensive preprocessing and modeling pipeline.

Another gap lies in the lack of robust feature selection strategies tailored to handle noisy or redundant features in HCV datasets, which can degrade model performance. Most prior works also favored complex black-box models like neural networks, which, although accurate, sacrifice interpretability, an essential requirement in medical decision-making.

Our research addresses these gaps by placing dataset balancing as the cornerstone of our methodology. We systematically incorporate advanced oversampling techniques with ensemble models like Extra Trees, ensuring fair representation of all fibrosis stages while maintaining high interpretability. Moreover, we adopt a rigorous feature selection and validation process to improve generalizability and reduce overfitting. By doing so, our study offers a balanced, interpretable, and clinically applicable framework that can reliably predict HCV stages across diverse patient populations.

CHAPTER III
METHODOLOGY

3.1 HEALTHCARE SYSTEM'S STRUCTURE

The healthcare system's structure for medical diagnostics is illustrated in Figure 3 which shows the process from the medical examination and at that point of data collection from the patients as an initial step in the process. The system that uses computer-aided diagnosis (CAD) to analyze and produce the results and receives the aggregated data. The healthcare provider receives these results and uses an XAI (Explainable Artificial Intelligence) framework to assist in the interpretation process. This framework helps keep the diagnostic process open and honest by making sure everyone knows what to expect at every step. It also gives doctors the tools they need to make good diagnoses using AI.

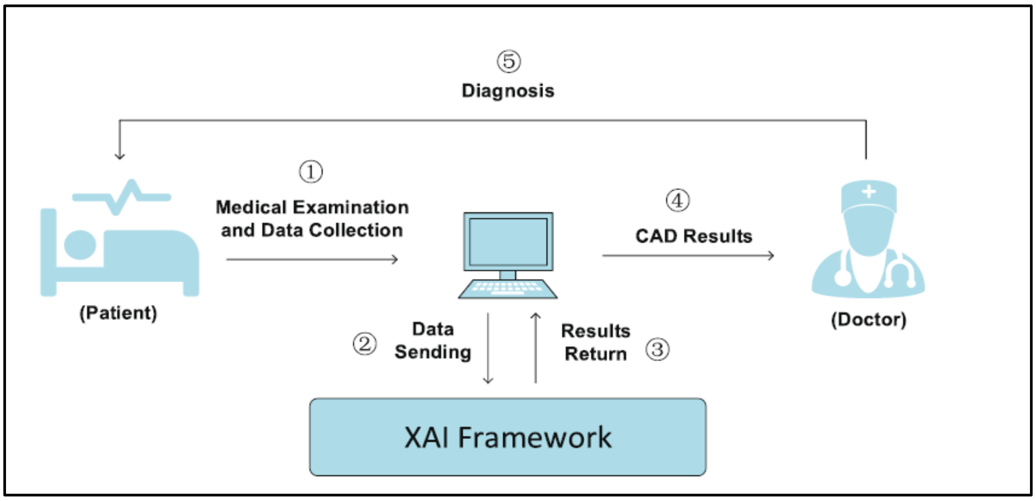


Figure 3: General healthcare system's structure [33]

3.2 SUPERVISED MACHINE LEARNING

One of the fundamentals of machine learning, supervised learning comprises training models with labelled datasets. The algorithm is able to understand the relationship between variables and predict outcomes for fresh, unknown data since the input data is connected to relevant output labels. SML frequently fixes regression and classification issues. Data is organized into certain groups by the model. It can identify diseases in medical records, locate objects in photos, and distinguishes spam from authentic communications. Regression concerns aim to forecast changing variables like stock prices, property values, and weather [45].

Evaluation of the success of supervised learning models often involves the utilization of a deferent measures, that vary according to the specifics of the task at hand. It is usual practice to use performance metrics as assessment criteria when dealing with classification challenges. MSE, RMSE, and R-squared values are commonly employed as metrics for evaluating regression model performance.   that could ensure models apply effectively to unseen data and avoid overfitting [46].

3.3 IMPORTANCE OF CLASS BALANCING IN MEDICAL DATASETS

Balancing the classes in medical data is crucial for ensuring that predictive modelling is equitable and precise in detecting illnesses and assessing risks. In the realm of medicine, statistical data is imbalanced, exhibiting a much higher frequency of occurrences in the majority class compared to the minority class. This discrepancy poses significant challenges for machine learning models, since they often exhibit a bias towards the majority class, which may hinder their ability to identify critical minority instances. The misclassification of high-risk individuals may result in significant repercussions, including postponed diagnosis and insufficient treatment [47].

Synthetic Minority Oversampling Technique (SMOTE) will be used to identify the predominant and often successful sampling method. The technique developed in 2002 has been used to address numerous difficulties related to imbalanced datasets [48]. It contrasts with random sampling methods by producing synthetic samples using the k-nearest neighbors of the analyzed samples, instead of replicating data from the minority class. This strategy largely emphasizes the development of synthetic data in contrast to competing approaches. The main goal is to produce more instances of minority classes by interpolating between various occurrences of an existing minority class [49].

3.4 PROPOSED METHOD

The proposed approach for forecasting HCV and liver-related illnesses comprises multiple organized processes to guarantee elevated prediction accuracy and model resilience, as seen in Figure 4. Initially, many publicly accessible datasets will be gathered and processed for machine learning analysis. Features will be chosen to eliminate unnecessary or duplicate attributes. The datasets will be partitioned into subgroups for training and evaluation. To rectify the class imbalance problem, SMOTE will be used to achieve equitable distributions. Classification models, such as the Extra Tree Classifier (ETC), will then be trained on the balanced training dataset. The trained models will ultimately be assessed on the test dataset using several metrics.

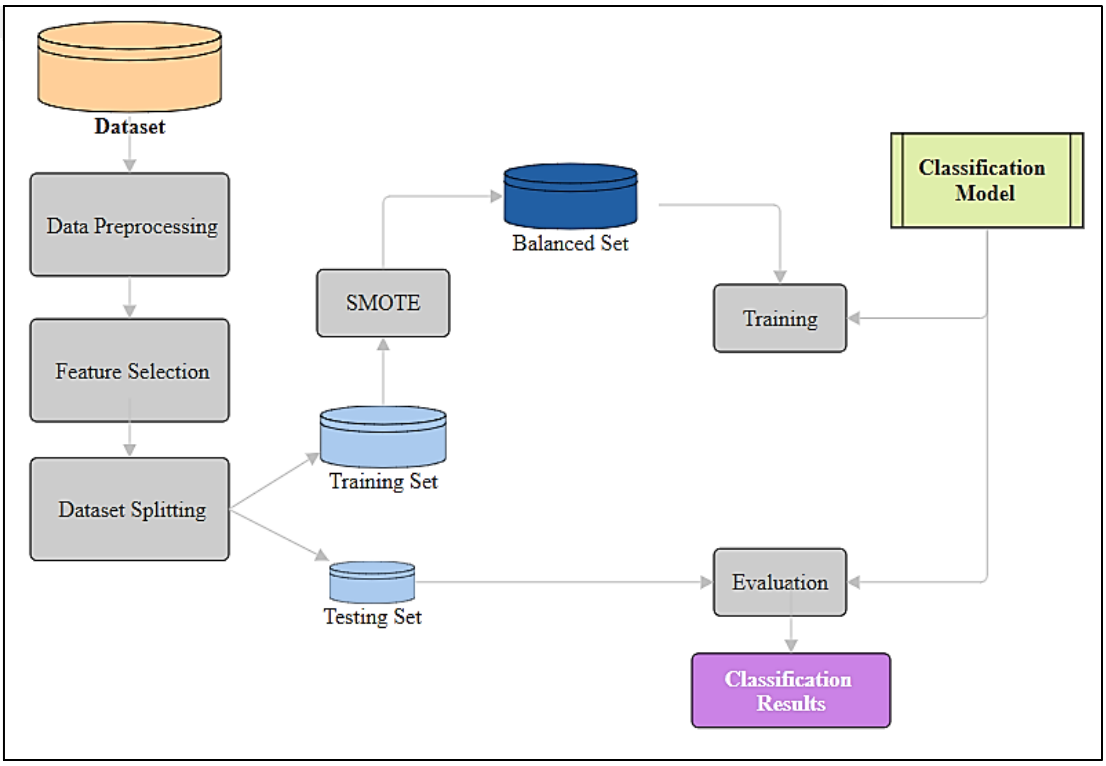


Figure 4: General methodology workflow

At the end of the process, the results are analyzed and the value scores of the features are shown. This lets us find the most important factors for finding hepatitis C virus and liver disease. The parts that follow will talk about each step.

3.5 DATASET

The exploration of medical data mining has recently emerged as a prominent subject of interest in the field of data analysis. A multitude of researchers have explored the development of sophisticated medical decision support systems aimed at assisting physicians [50]. The healthcare data classification process aids clinical practitioners in diagnosing and treating medical disorders and ailments in patients. It functions as a decision support system (DSS), significantly enhancing the quality of healthcare [51]. The implementation of this study consists of Three datasets first one titled with" HCV-EGY-DATA" and second one titled with "HCV dataset" and third one titled with "ILPD" datasets are free access on the UCI Repository.

3.5.1 HCV-EGY Dataset

Donated on 9/29/2019**,** the HCV-EGY-DATA dataset, licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, and accessible through the UCI Machine Learning Repository, DOI **10.24432/C5989V** permitting sharing and adaptation with proper credit. The dataset comprises information on 1,385 Egyptian Hepatitis C therapy recipients, encompassing 14 attributes collected over an 18-month period [66]. There is great value in this dataset. for developing predictive models and conducting analyses on HCV treatment outcomes providing key insights into disease progression and management within the Egyptian population [20]. The dataset contains 29 attributes described in Table 3. HCV-EGY dataset has no missing values but it's suffering from the class imbalanced distribution which should leading to use SMOTE for the balancing class distribution.

Table 3: Attributes description in the HCV-EGY dataset

No	Description	No	Description
1	Age	16	ALT 1 (1week) - Alanine transaminase ratio 1 week
2	Gender	17	ALT 4 (4week) - Alanine transaminase ratio 4 weeks
3	BMI (Body Mass Index)	18	ALT 12(12week) - Alanine transaminase ratio 12 weeks
4	Fever	19	ALT 24(24 week) - Alanine transaminase ratio 24 weeks
5	Nausea/Vomiting	20	ALT 36(36week) - Alanine transaminase ratio 36 weeks
6	Headache	21	ALT 48(48week) - Alanine transaminase ratio 48 weeks
7	Diarrhea	22	ALT (after 24week) – Alanine transaminase ratio after 24 weeks
8	Fatigue & generalized bone ache	23	RNA Base

Table 3 Cont.

No	Description	No	Description
9	Jaundice	24	RNA 4
10	Epigastric pain	25	RNA 12
11	WBC (White blood cells)	26	RNA EOT (RNA End-of-Treatment)
12	RBC (Red blood cells)	27	RNA EF (RNA Elongation Factor)
13	HGB (Hemoglobin)	28	Baseline Histological Grading (Target)
14	Plat (Platelets)	29	Baseline Histological Staging (Class labels)
15	AST 1 - Aspartate transaminase ratio		

The dataset exhibits a complex and imbalanced distribution across 14 grading intervals of the Baseline Histological Grading target variable, ranging from 3.00 to 16.00. Some grading ranges, such as 3.00–4.30 (183 cases) and 14.70–16.00 (225 cases), are highly represented, while others like 4.30–5.60 (93 cases) and 9.50–10.80 (87 cases) are underrepresented. This uneven distribution can bias machine learning models towards the dominant grades, reducing sensitivity for minority grading intervals. This problem must be addressed so that the classifier learns equally from all levels of classification and improves generalization.

3.5.2 HCV Dataset

Donated on 6/9/2020, An international Creative Commons Attribution 4.0 (CC BY 4.0) license governs the use of this dataset. Datasets for any purpose can be shared and modified in this way, provided proper attribution is given. DOI: 10.24432/C5D612. The medical information contained within this dataset includes demographic information related to individuals, including blood donors and patients diagnosed with hepatitis and liver cirrhosis. The dataset consists of 615 entries and 13 attributes, which include age, sex, and several biochemical markers The "category" column classifies individuals into distinct groups: blood donors, suspected blood donors, patients with hepatitis, patients with liver cirrhosis, and liver cirrhosis patients. This dataset could be important for research in the medical fields, especially in the study of liver diseases and the associated biomarkers. Missing data (NA values) highlights the need for data preprocessing before analysis. Addressing these gaps is essential for accurate and reliable results. This dataset offers a strong base for exploring relationships among biochemical indicators and liver health. Its structure and range of markers make it ideal for identifying correlations and patterns [67]. Table 4 details the five main classification categories and their sample sizes, which is crucial for interpreting analytical results and drawing valid inferences about the links between biochemical indicators and liver health conditions. This information helps contextualize the data and understand its scope and limitations, ensuring the validity of conclusions [52].

Table 4: Attributes description in the HCV dataset

Type	Field Name	Field Description	Normal value range
Continuous feature	Age	-	F and M
	ALB	Albumin	35–55 g/L
	ALP	Alkaline phosphatase	0–40 U/L
	ALT	Glutamicpyruvic transaminase	M: 5–40 U/L F: 5–35 U/L
	AST	Glutamic oxaloacetic transaminase	0–40 U/L
	BIL	Bilirubin	5.10–19 μmol/L
Classification feature	CHE	Serum cholinesterase	4.3–10.5 U/L
	CHOL	Total cholesterol	2.83–5.18 mmol/L
	CREA	Creatinine substance	M: 50–110 μmol/L F: 40–100 μmol/L
	GGT	Glutamyl transpeptidase	3–50 U/L
	PROT	Total protein	20–80 mg/L
	SEX	-	F and M

The dataset exhibits a significant imbalance in class distribution. Specifically, the majority class, “Blood Donor” (labeled as 0), constitutes 87% of the dataset, while the “Cirrhosis” class (labeled as 3) accounts for only 5%. The remaining classes, including “Suspected Blood Donor,” “Hepatitis,” and “Liver Cirrhosis,” collectively make up just 8% of the entries. This disproportionate representation can bias the classifier toward the dominant class, reducing its ability to detect minority categories effectively.

3.5.3 Indian Liver Patient Dataset (ILPD)

One such publicly accessible dataset is the Indian Liver Patient Dataset (ILPD), which is housed in the UCI Machine Learning Repository. donated on May 20, 2012, by the Andhra Pradesh Institute of Medical Sciences and Teaching Hospitals, India This dataset consists of medical diagnostic data collected from 583 [68]. This dataset was created to develop and test machine learning models capable of classifying and predicting liver-related disorders. Each case in this dataset contains 10 biological and demographic characteristics, the most important of which are age, gender, bilirubin levels, albumin levels, and liver enzyme levels including SGPT, SGOT, and ALP, all of which are important indicators of liver function efficiency. A binary categorization that indicates whether the patient has liver disease (1) or not (2) is the final feature. The researchers in the fields of medical informatics, and the diseases predictions center, and health-related data analytics might use the ILPD dataset. This dataset may be advantageous for feature selection research and also algorithm validation in biological data science because of its clinical importance and diverse feature set. Table 5 shows the dataset's feature structure.

Table 5: Attributes description in the ILPD dataset

Variable Name	Type	Description	Missing Values
Age	Integer	Age	no
Gender	Categorical	Gender	no
TB	Real Number	Total Bilirubin	no
DB	Real Number	Direct Bilirubin	no
Alkphos	Integer	Alkaline Phosphotase	no
Sgpt	Integer	Alamine Aminotransferase	no
Sgot	Integer	Aspartate Aminotransferase	no
TB	Real Number	Total Proteins	no
ALB	Real Number	Albumin	no
A/G Ratio	Real Number	Albumin and Globulin Ratio	4
Selector	Binary	Absence or presence of disease	no

The dataset also showed a class imbalance, including 416 (71.3%) with liver diseases and 167 (28.7%) healthy individuals. This imbalance poses challenges such as biasing model predictions toward majority classes and poor generalization to minority classes, which must be addressed.

3.6 DATA PREPROCESSING

3.6.1 HCV-EGY Data Preprocessing

Several preprocessing processes have been applied to this dataset to guarantee data quality and model efficacy. To begin with, a Pandas DataFrame was used to load the dataset. Then, features were chosen according to their significance. The last 19 columns before the target variable, " Baseline histological Grading," examined carefully. To make the evaluation of the model, the dataset has been divided to 80% for training and 20% for testing. To address and avoid the issue of class imbalance, SMOTE was applied to the training data. The resulting method of the synthetic samples for the minority class enhances the model's generalization capability. An Extra Trees has been trained on the resampled dataset to identify the most importance features, after the dataset was balanced, were ordered according to their significance scores. The feature importance was used to bring clarity to the most significant predictors.

3.6.2 HCV Data Preprocessing

The process of preparing for this dataset starts from deal with missing values, and also categorical variables should be encoded, and class imbalance handled by SMOTE, ln order to enhance the model's performance, and feature selection has been implemented. The dataset was first loaded and examined to check if the missing values existed. the numerical features with missing values have been calculated using the mean method, while categorical features, like "gender," have been filled with the mean value. Categorical variables have been encoded by using the LabelEncoder to convert them into numerical representations suitable for the models. The dataset was also divided into features and a target variable, with the removal of the unnecessary columns. The SMOTE technique was used to correct and deal with the class imbalance the dataset was suffering from. Keeping the class proportions constant, the dataset was divided into 80% for training and 20% for testing. The Category column is set as the target variable (y), which represents the class labels to be predicted. The target variable is encoded using LabelEncoder for machine learning compatibility. The Category column has five unique classes. Then an extra tree classifier was trained using optimized hyperparameters to reduce overfitting. The importance of the features has been determined by their contribution to the impurities reducing across all trees.

3.6.3 ILPD Data Preprocessing

The dataset contains 582 records with 11 features linked to the diagnosis of the liver diseases of Indian people. This dataset went through a number of preprocessing processes for maintain the data quality and to improve the model performance. Then the dataset is loaded to test the data types and discover if there are any missing values that exist. This was particularly important in the "Albumin to Globulin Ratio" column, which contained some of the empty entries. Categorical values like "gender" were converted into binary numbers 0 and 1. A feature importance analysis was carried out to determine which features were most likely to have an impact on the categorization. The results demonstrated that the most critical aspects are bilirubin levels and liver enzymes. Using oversampling techniques such as SMOTE and RandomOverSampler, to handle the issue of class imbalance in the target variable ("Selector"), it would be prudent to train the models on a balanced dataset to attain strong and unique performance.

3.7 FEATURE SELECTION

The goal of feature selection, a basic preprocessing step in data analysis and machine learning, is to isolate and keep the most important features while removing any superfluous or unimportant ones. This process is crucial in high-dimensional datasets, where excessive Features may result in overfitting, heightened computing complexity, and diminished model interpretability [53]. Feature selection is widely applied across various domains. In biomedical research, it aids in identifying significant biomarkers from genomic data, improving disease prediction and classification models [54]. Image processing and remote sensing also benefit from feature selection by reducing the computational cost of analyzing large-scale image datasets [55]. Feature selection is to pick a group of variables from the input that accurately describes the data while reducing the impact of noise or factors that aren't relevant, ensuring strong predictive performance [56].

3.7.1 Feature Selection for Classification

The training phase of classification is where feature selection has the most impact. Feature selection for classification works by identifying a subset of features after feature generation and then processing the data using these features in the learning algorithm, as opposed to just inputting the whole set of features. There are two types of feature selection algorithms: those that work independently of learning (filter models) and those that iteratively use learning algorithm performance to evaluate feature quality (wrapper models). In order to make predictions, a classifier is trained on the features that were chosen. Finding the most minimal group of features using the following criteria is a common goal of feature selection in classification: The accuracy of the classification has not altered much. this results in a conserved class distribution that is based entirely on the values of the characteristics that were chosen [57].

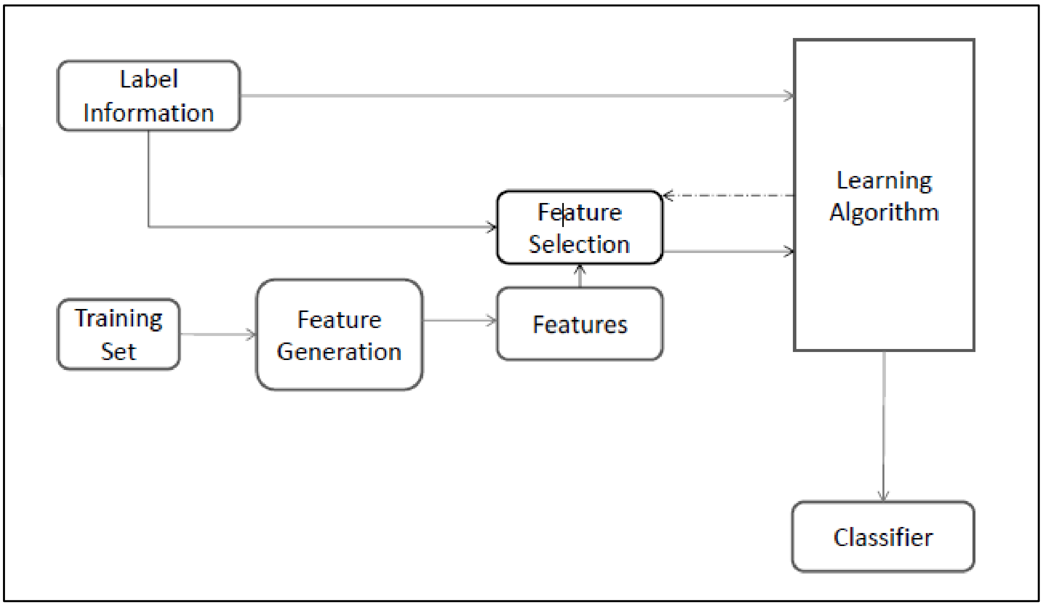


Figure 5: General framework of Feature selection for classification [57]

3.7.2 Feature Importance Analysis

In order to determine which features were most important for predicting the target variable, an Extra Trees Classifier was employed. The target variable has been eliminated during model training from the HCV-EGY and HCV datasets. and each attribute's worth was determined by how much it helped with impurity reduction in decision trees. The important ratings were extracted using the feature_importances_attribute of the Extra Trees Classifier. Higher scores indicate that the ratings were more relevant to the classification task.

In ILPD dataset, feature selection was performed using a cross-validated Random Forest, which aggregated feature importance scores across five folds, identifying Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Age, Total Bilirubin, Albumin, and Total Proteins as the top seven predictors. These features were extracted to reduce dimensionality and improve model performance.

This study provides valuable insights into the key features that influence the model's predictions, making it easier to identify the most significant components of classification.

3.8 DATASET SPLITTING

The dataset was divided into training and testing subsets using an 80:20 split to ensure robust model evaluation. The training set was used for model fitting, while the testing set was reserved for assessing the model's performance on unseen data.

3.9 SMOTE

This method was suggested utilizing the oversampling technique to augment the sample size of minority classes. SMOTE operates on the principle of creating synthetic samples by randomly interpolating between a minority class sample and its nearest neighbors within the same class. In SMOTE, the distance between the minority class and the k-nearest neighbors is computed. The outcome derived from the sample is incorporated into the training dataset after the multiplication of the differences by a random value ranging from 0 to 1. Consequently, SMOTE operates by randomly augmenting existing instances with additional points derived from neighborhood ties. Consequently, SMOTE possesses the capability to augment the sample size of the minority class, akin to random oversampling. Nonetheless, it does not produce an identical instance. It produces a novel sample from the current samples, hence mitigating the overfitting issue associated with the oversampling technique to a certain degree. To address class imbalance in ILPD, both SMOTE and RandomOverSampler techniques were applied to the training set before model training.

3.10 CLASSIFICATION MODEL

Classification is a fundamental technique used in data analysis, which can be used to categorize data into predefined groups based on features. They can be classified broadly into supervised and unsupervised classification, with different classification methods [58]. Despite its effectiveness, classification faces challenges such as handling missing data, computational complexity and bias, necessitating ongoing improvements in algorithms and data preprocessing [59]

The core target of data set classification is to sort data according to its characteristics. The purpose is to gather information in order to establish relationships between them. Data for training and data for testing are separated from this whole dataset. By randomization, 80% of the dataset is designated for training purposes, while 20% is set aside for testing. When building a learning model, the data used for training is known as the training data. In order to assess the model, the test data is utilized. Machine learning is the discipline that allows computers to acquire knowledge from data through algorithms. This cannot be accomplished if machines are not allowed to acquire experience and use the knowledge they have obtained. We can alleviate numerous jobs for humans through machine learning. This sector is crucial due to its numerous advantages, such as identifying relationships in extensive datasets, managing vast quantities of data that are impractical for human processing in a short timeframe, assisting expert decision-making (e.g., a doctor's diagnosis), and processing image data. Supervised learning and unsupervised learning are the two categories of learning. In recent years, machine learning techniques have been applied to health data analysis. These methods utilize class labels to denote data across many categories. There are two primary theoretical frameworks within machine learning. computer learning is carried out in an unsupervised learning approach without imparting any class knowledge to the computer. The goal of the algorithm utilized in this method is to have the machine classify based on the similarity criteria. Unsupervised learning is the sole option when the data class is unknown.

To train a model to the results of the data that it has never encountered previously, supervised learning relies on providing the machine with information about data it has never seen before. Another way of looking at it is that the machine learns to understand the problem by first being trained using an algorithm employing data that includes the class information. The next step could be training the algorithm to perfectly classify similar data in the future. This is accomplished by initially splitting the dataset into two subsets. During one stage, known as the learning phase, one uses these two groups, while in the other, known as the test phase, it uses the other. Because it is feasible to end up with distinct groups as a result of dividing the courses, several approaches are taken to ensure that there is no discernible difference between the two sets of students.

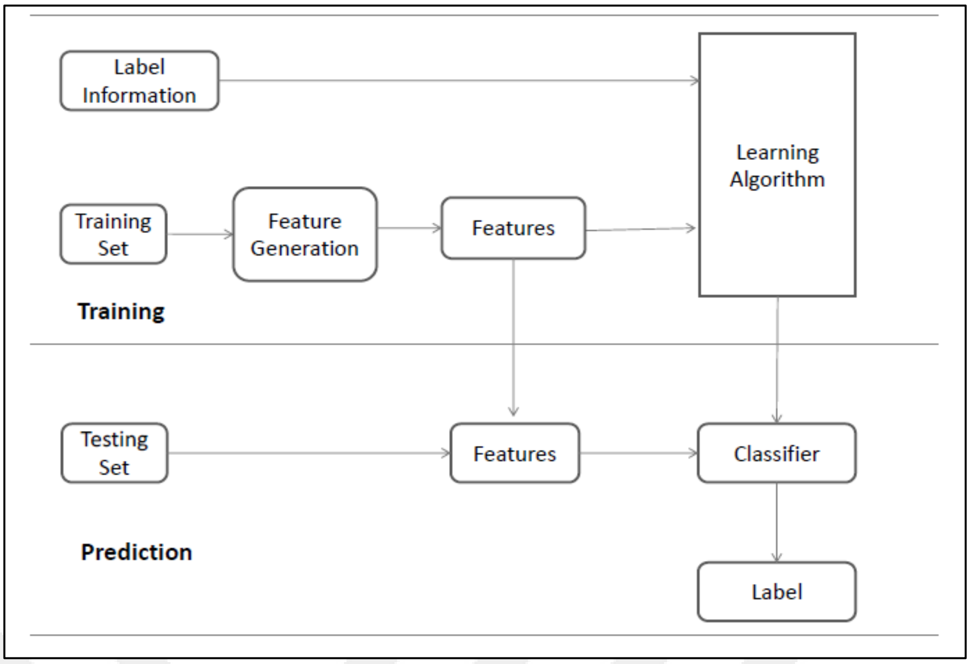


Figure 6: A General process of data classification [57]

Extra Tree Classifier (ETC) integrates the forecasts of numerous DTs using a series of ML algorithms. The widely-used RF algorithm is extremely comparable. Despite employing a less complex technique to generate group decision trees, it typically outperforms the RF approach. Decision or regression trees that do not undergo standard top-down pruning are produced by the additional tree classifier. It uses the whole learning example to create trees and separates nodes by randomly setting thresholds, which are two important differences from existing tree based ensemble approaches [60]. The additional trees classifier divides tree nodes based on random cut points. It also creates trees from the entire training dataset, as opposed to bootstrapping, which would necessitate recreating it.

In HCV-EGY dataset, the ETC was utilized as the core model for classification due to its efficiency in handling high-dimensional data and its ability to compute feature importance. The ETC was configured with 100 estimators (n_estimators=100) to build an ensemble of decision trees for robust predictions. The random_state=42 parameter ensured reproducibility of results by controlling randomness in the model.

In HCV dataset, the ETC was configured with specific parameters to optimize performance and prevent overfitting. The max_depth=10 parameter restricted the depth of the trees to control model complexity and improve generalization. The min_samples_split=10 ensured that a node must have at least 10 samples to be split further, reducing the likelihood of creating overly specific rules for small data subsets.

Similarly, min_samples_leaf=5 required each leaf node to contain at least five samples, which helps smooth out predictions for rare classes. The random_state=42 parameter was set for reproducibility of results.

In ILPD dataset, several ensemble-based models implemented on this dataset were carefully configured with parameter settings aimed at achieving optimal performance and preventing overfitting. For the Random Forest and Extra Trees Classifier, the number of estimators (trees) was set to 150 to ensure sufficient diversity and stability in predictions. The maximum depth of trees was kept at a moderate level (max_depth=10) to avoid excessive complexity, while min_samples_split=10 and min_samples_leaf=5 was used to control tree growth and improve generalization. For XGBoost, the parameter eval_metric was set to logloss to align with the binary classification task, and regularization was applied to reduce overfitting. Similarly, LightGBM was configured with default parameters and tuned during cross-validation to balance speed and accuracy. These parameter settings collectively ensured that the models could handle class imbalance, learn effectively from resampled data, and deliver high predictive performance in both training and testing phases.

3.11 TRAINING

During the training phase, the classifier was trained on the SMOTE-balanced training set. The ETC aggregated predictions from all trees using majority voting, effectively reducing variance and overfitting. This training approach allowed the model to learn complex relationships between selected features and the target variable.

In ILPD dataset, a 5-fold stratified cross-validation strategy was applied, maintaining the proportion of classes in each fold and ensuring balanced performance assessment. The ROC-AUC score was chosen as the primary evaluation metric due to its ability to measure the classifier’s discriminative power across thresholds, especially important for imbalanced datasets.

3.12 EVALUATION

When testing the model on test sets, several evaluation criteria were used.

3.12.1 Confusion Matrix

A confusion matrix is a chart that defines the correlation between the algorithm's predictions and the actual outcomes of the classification executed by the algorithm. This matrix serves to assess performance, as the values contained within it are employed for this purpose. The actual scenario is represented by either the row or the column, while the estimation is provided by the alternative [61].

- True positive (TP): Data accurately identified as positive.
- True negative (TN): Data accurately identified as negative.
- False Positive (FP) Data inaccurately classified as positive.
- False Negative (FN): Data inaccurately labelled as negative.

Table 6: Confusion Matrix Formula

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

3.12.2 Accuracy

Accuracy is the paramount metric of categorization performance. The ratio of true positives (individuals correctly recognized as patients by the algorithm) and true negatives (individuals accurately identified as non-patients) is compared to the total values. This aspect of the issue pertains to genuine positives, projected positives, and actual state positives. For instance, it aids in identifying the individual by classifying them as a patient.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.1}$$

3.12.3 Precision

The number of accurately anticipated positive classes that are truly positive is called precision. In other words, it is the chance that someone who has a positive diagnostic test result has the condition.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

3.12.4 F1-score

The F1 score is computed by averaging the recall and precision values. This allows for the simultaneous evaluation of the two metrics [62].

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3.3}$$

3.12.5 Recall

The ratio of correctly identified positive data to the total positive data, expressed as a percentage.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.4}$$

3.12.6 Classification Report

A classification algorithm's performance can be assessed with this tool. by providing a detailed breakdown of various metrics. It includes precision and recall as well as F1-score and support for each class. Generated for each model to visualize performance.

CHAPTER IV
EXPERIMENTS AND RESULT

This section will cover the analysis of the results of all datasets individually and main finding in this study, also will distinguish between the results obtained from these datasets. as will discuss these finding with the previous studies implementation and results.

The implementation and evaluation of all experiments in this proposed work has been performed on custom configured workstation operating on Microsoft Windows 11 (Version 10.0.22631). The computational infrastructure consisted of a 12-core Intel processor (Family 6, Model 165, Stepping 2) with a base clock speed of 2.59 GHz, supported by 16 GB of DDR4 memory, of which approximately 64% was utilized during peak execution periods. The graphical processing capabilities included both Intel UHD Graphics and a dedicated NVIDIA GeForce GTX 1650 Ti, facilitating efficient parallel computation for model training and data visualization. Storage resources were allocated across three NTFS-formatted drives (C: D: and E:) with an aggregate capacity exceeding 1.4 TB, ensuring sufficient space for datasets and intermediate outputs.

The software stack was based on Python, incorporating a range of specialized libraries for data science and machine learning. These included NumPy and Pandas for numerical operations and structured data manipulation, Matplotlib and Seaborn for advanced visualization, and Scikit-learn for preprocessing, feature scaling (e.g., RobustScaler, PowerTransformer), and model training. Ensemble-based learning algorithms, including Extra Trees, were employed for classification tasks. Class imbalance issues were tackled using the SMOTE and RandomOverSampler, both of which were incorporated using Imbalanced-learn (imblearn) pipelines.

4.1 HCV-EGY DATASET

Accuracy, precision, recall, F1 score, and mean squared error were among the numerous performance metrics used to evaluate the extra tree classifier. Table 7 displays the 98% accuracy rate of the model, The precision and recall were flawless at a rate of 98%, which significantly indicates that there have been zero instances of false positives or negatives, given the model's exceptional success in detecting positive cases. F1-score 98% shows that accuracy and recall are well balanced. A common aspect of the classification issue is this: There was an extremely low mean absolute error rate.

Table 7: Classification report of ETC on HCV-EGY dataset

Class	Precision	Recall	F1-score	Support
3	1.00	1.00	1.00	18
4	0.95	1.00	0.97	19
5	0.95	0.95	0.95	22
6	1.00	0.95	0.97	20
7	1.00	1.00	1.00	16
8	1.00	1.00	1.00	20
9	1.00	1.00	1.00	20
10	1.00	0.92	0.96	13
11	0.95	1.00	0.98	20
12	1.00	0.95	0.97	19
13	0.96	0.96	0.96	24
14	0.95	1.00	0.98	20
15	1.00	1.00	1.00	28
16	1.00	1.00	1.00	18
Accuracy			0.98	277
Macro avg	0.98	0.98	0.98	277
Wighted avg	0.98	0.98	0.98	277

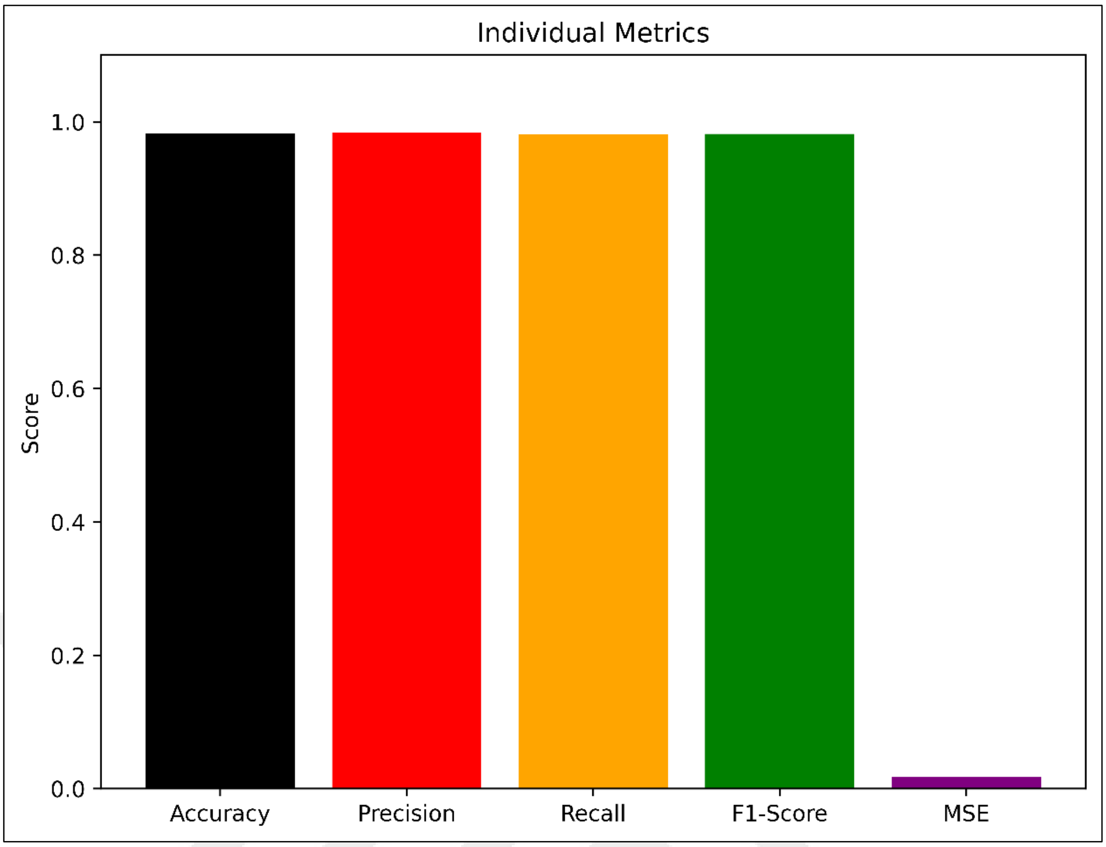
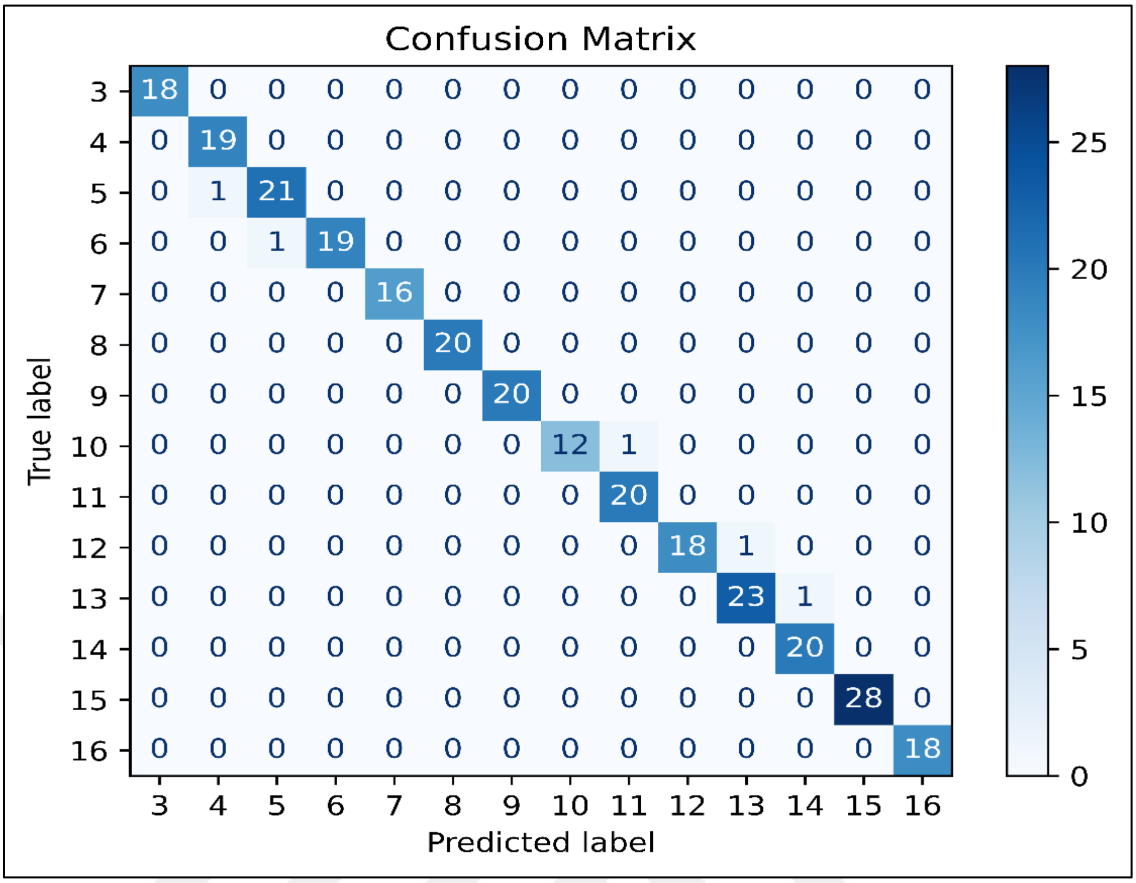


Figure 7: Performance metrics of ETC on HCV-EGY dataset

Figure 8's confusion matrix indicates that instances were accurately classified into both the true positive and true negative categories. The matrix shows only a few misclassifications, indicating that the classifier was highly efficient in distinguishing between the classes. Accurate positives, accurate negatives, erroneous positives, and their respective breakdowns, the model's exceptional predictive capacity is further demonstrated by its low false negative rate.



	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3	18	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	19	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	21	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	19	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	16	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	20	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	20	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	12	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	20	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	18	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	23	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	20	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	28	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	18

Figure 8: Confusion matrix of ETC on HCV-EGY dataset

In Figure 9, feature importance analysis highlighted the most significant predictors for the classification task. Baseline histological grading was the most crucial feature contributing heavily to the forecasts made by the model, with a significance level of around 0.40. Other key features included RBC, RNA 4, RNA Base, and ALT 36, each showing notable importance in the classification decision. These findings could emphasize the essential role of clinical and test result features in accurately predicting the target variable.

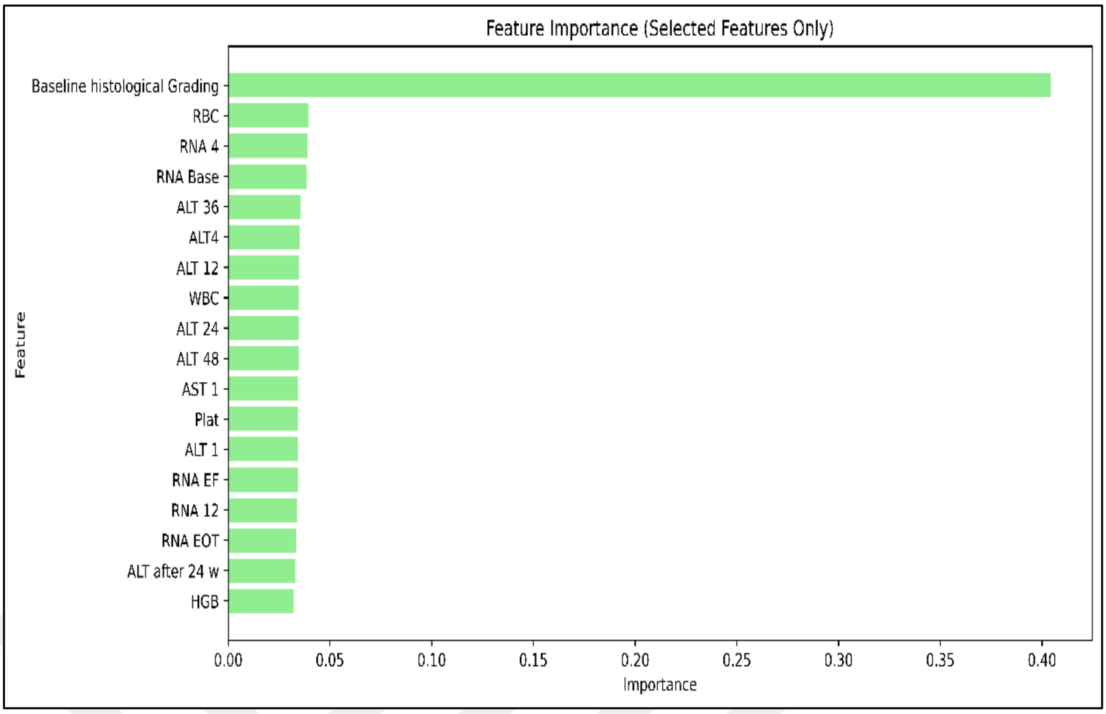


Figure 9: Feature importance for HCV-EGY dataset

4.2 HCV DATASET

Bar chart in Figure 10, and Table 8 demonstrates the major performance metrics of the classification model in Figure 16, including Accuracy, Precision, Recall, and F1-score. All metrics are close to 98%, indicating strong predictive performance. The model is performing effectively in identifying the majority of the positive samples. Precision and recall, together with the F1-score, remain strong, indicating that the model is reliable and effective for classification.

Table 8: Classification report of ETC on HCV dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
0=Blood Donor	0.99	0.99	0.99	107
0s=suspect Blood Donor	0.99	1.00	1..00	106
1=Hepatitis	0.99	0.98	0.99	107
2=Fibrosis	0.99	0.99	0.97	107
3=Cirrhosis	1.00	0.95	0.98	106
Accuracy			0.98	533
Macro avg	0.98	0.98	0.98	533
Weighted avg	0.98	0.98	0.98	533

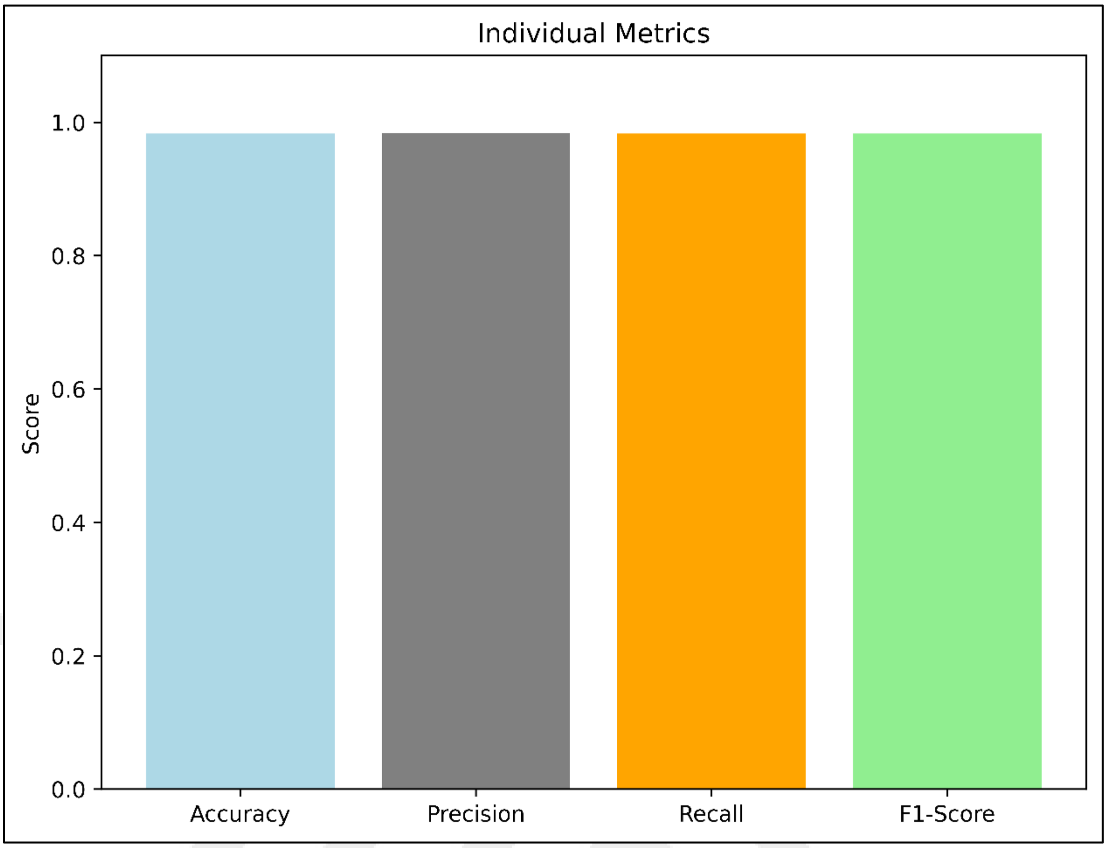


Figure 10: Performance metrics of ETC on HCV dataset

Figure 11 of confusion matrix evaluates the classification model's accuracy by comparing true labels (rows) with predicted labels (columns).

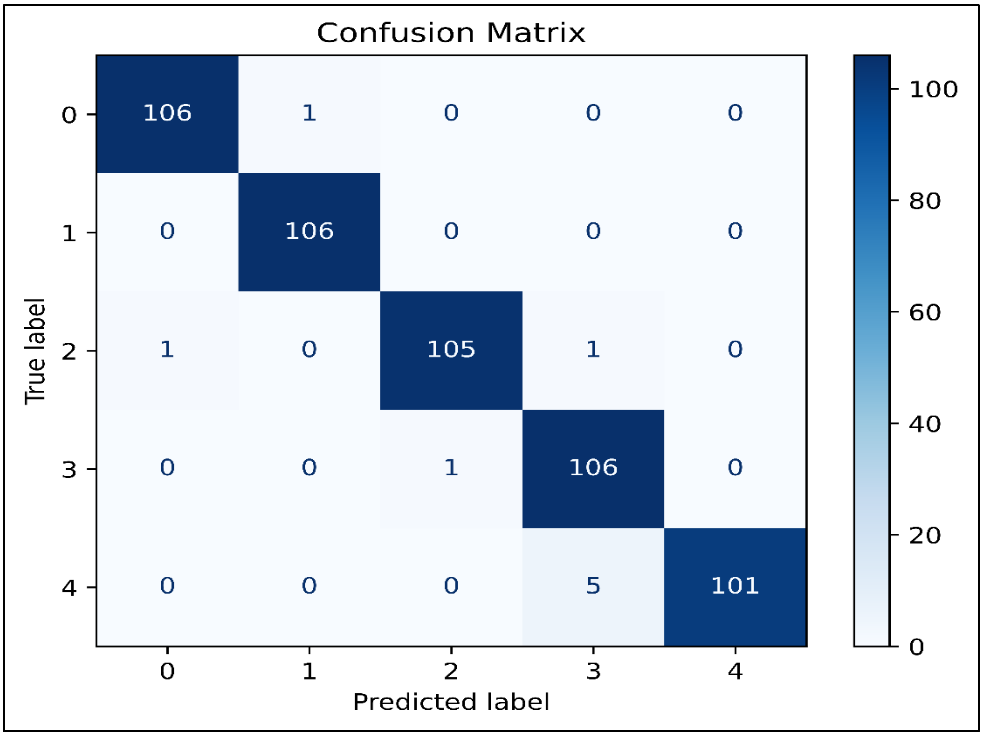


Figure 11: Confusion matrix of ETC on HCV dataset

The bar chart of feature importance of the HCV data in Figure 12 shows the most significant features in HCV prognosis prediction. Total protein (PROT), cholinesterase (CHE), and albumin (ALB) are among the most influential markers, highlighting the role of liver function and metabolic indicators in disease progression.

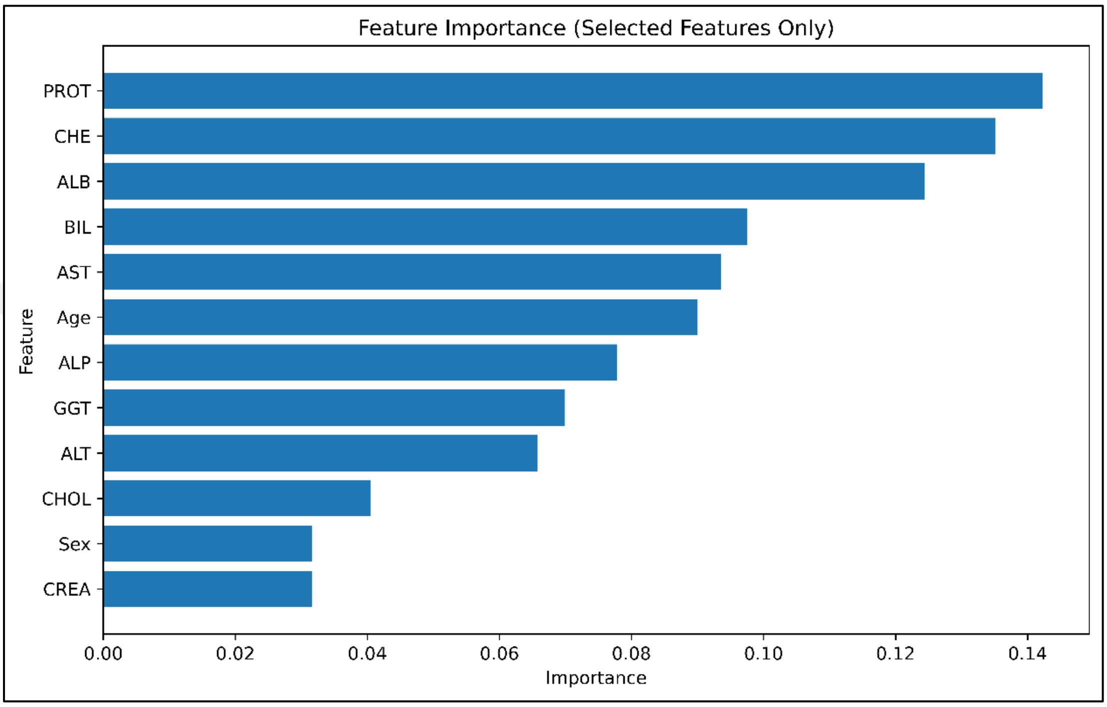


Figure 12: Feature importance for HCV dataset

4.3 ILPD DATASET

Table 9 presents the classification report, indicating a balanced performance across both classes (0 and 1), each comprising 83 instances. This model achieved an accuracy 95%, in both groups, the F1-scores of 0.95 and 0.94 indicate strong and consistent predictive capability.

Table 9: Classification report of ETC on ILPD dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.92	0.98	0.95	83
1	0.97	0.92	0.94	83
Accuracy			0.95	166
Macro avg	0.95	0.95	0.95	166
Weighted avg	0.95	0.95	0.95	166

Figure 13 presents a bar chart that illustrates consistent performance across all measures. Each performance metric is at 0.95. This suggests a robust and dependable model for data classification.

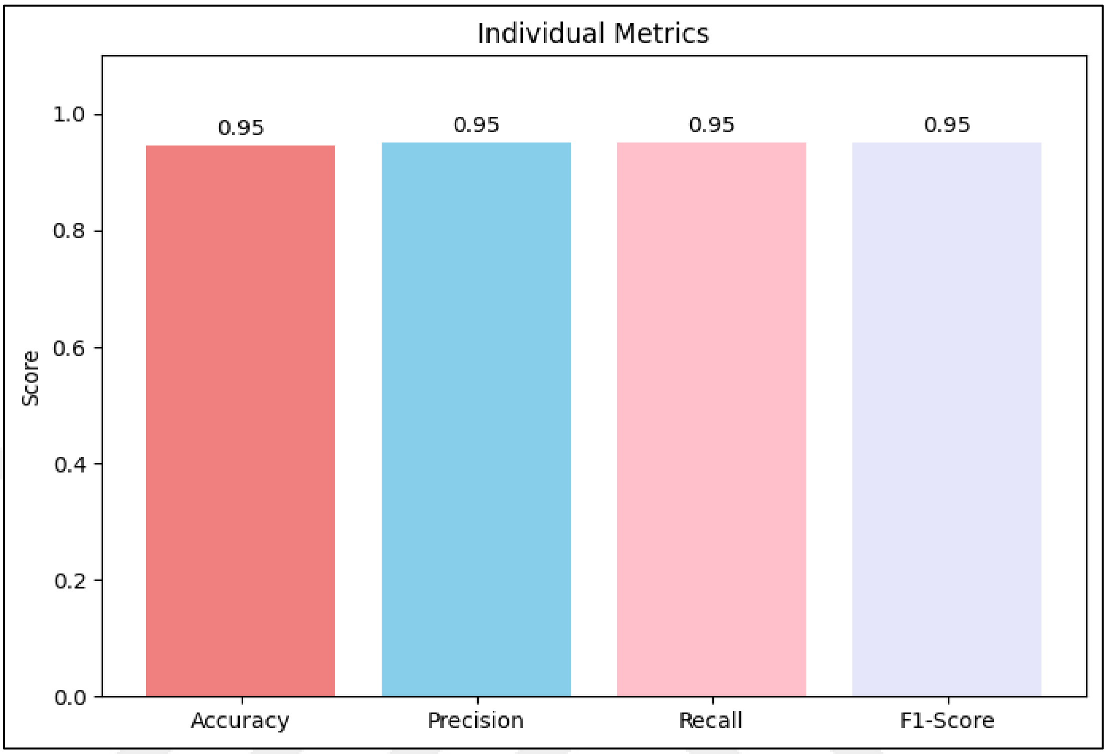


Figure 13: Performance metrics of ETC on ILPD dataset

Figure 14 of the confusion matrix shows that the ExtraTrees model correctly classified 81 out of 83 Class 0 samples and 76 out of 83 Class 1 samples. There were 2 false positives and 7 false negatives.

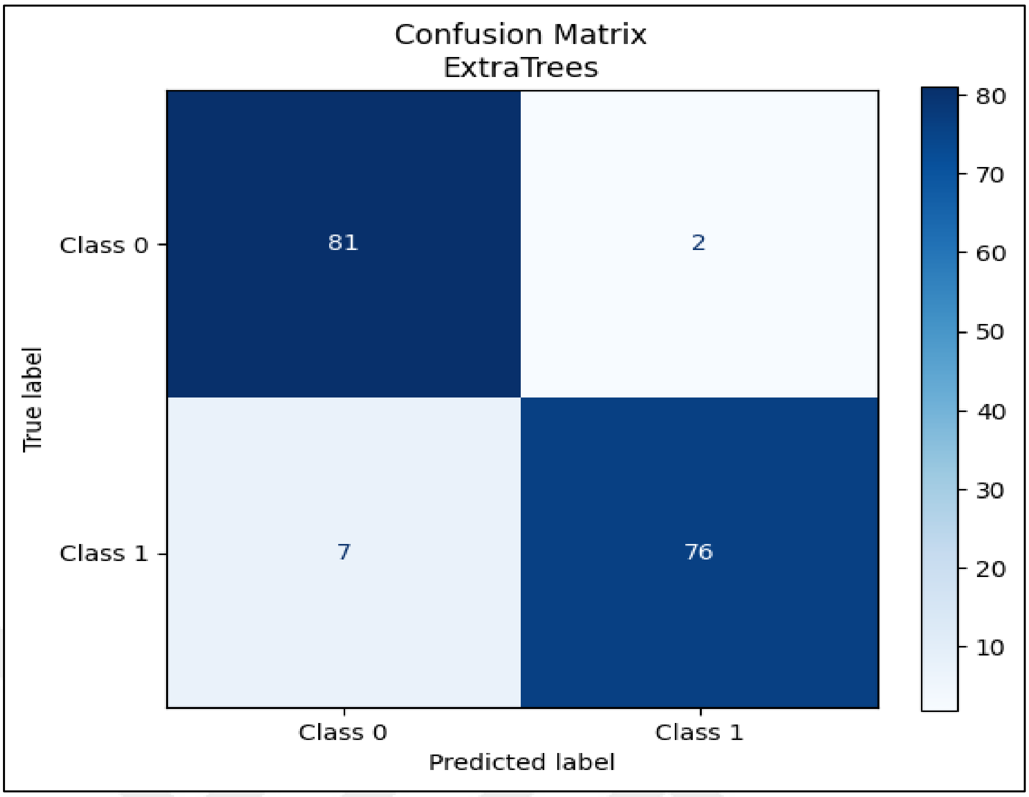


Figure 14: Confusion matrix of ETC on ILPD dataset

Figure 15 of feature importance chart shows that Total Bilirubin is the most influential factor in liver disease prediction. Other significant contributors include liver enzymes like Aspartate and Alkaline Aminotransferases, indicating their strong diagnostic value.

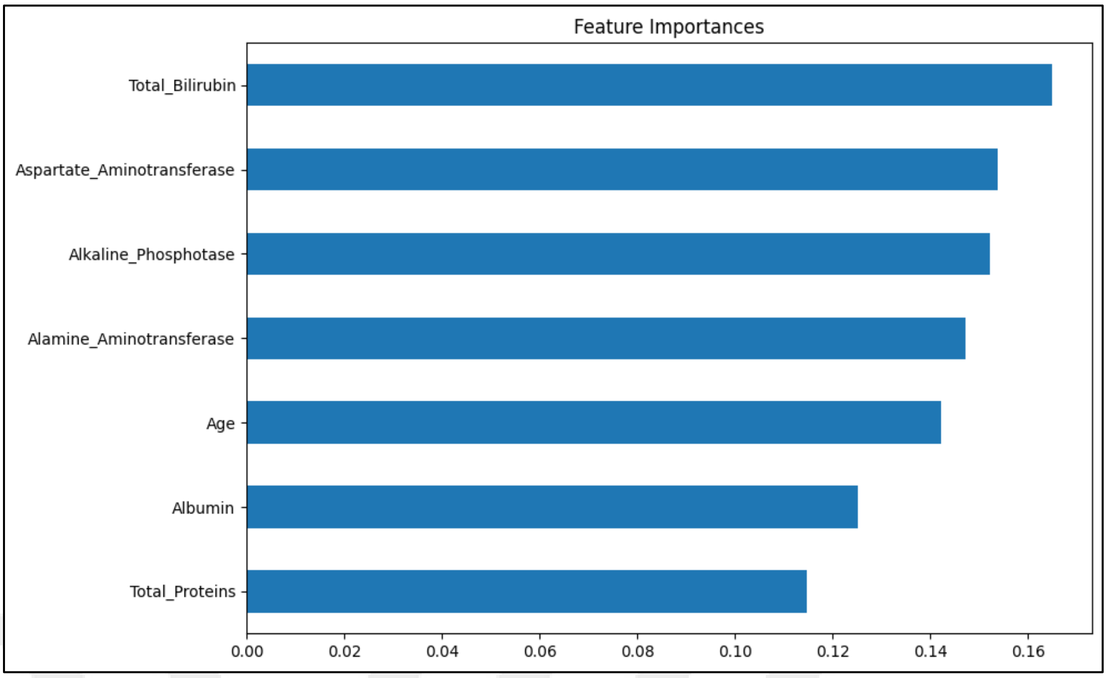


Figure 15: Feature importance for ILPD dataset

4.4 ANALYSIS OF RESULT

The model shows that it can classify well on the HCV-EGY dataset, with an accuracy of 98% and very few errors. The confusion matrix reveals a strong diagonal dominance, which means that the predictions are quite accurate with just a few small mistakes in few classes. The classification report shows that the recall, precision, and F1-scores are all the same, and that most classes got scores that were almost perfect. The bar chart analysis backs up the model's strength by demonstrating that all of the assessment metrics are near to 1.00 and that the Mean Squared Error (MSE) is low. According to feature significance analysis, "Baseline Histological Grading" is the most important aspect, whereas other characteristics are less important. The model is generally well-balanced, accurate, and mostly based on histology grading.

The confusion matrix for the HCV dataset demonstrates strong diagonal dominance, indicating that the classifier has performed well in correctly predicting classes. There are minimal misclassifications, with only a few instances of incorrect predictions, such as Class 4 being misclassified as Class 3 in five cases. The most consistent classifications are observed for Blood Donor (Class 0), Suspect Blood Donor (Class 0s), and Cirrhosis (Class 3), with nearly perfect classification.

The classification report for this dataset highlights an overall accuracy of 98%, which is similar to the HCV-EGY dataset. recall as well as accuracy and F1 scores were observed across all categories, confirming that the performance was balanced.

Liver cirrhosis (category 2) and liver cirrhosis (category 3) show a decrease in recall, ranging between 0.95 and 0.99, which explains that some cases of these conditions were misclassified into other categories. The overall and weighted averages of 0.98 can also confirm that the model works consistently across all categories.

The bar chart of the evaluation metrics shows these results, where there was an increase in performance metrics, which indicators of strong classification performance. The analysis of the line graph for performance metrics by categories shows that the suspected blood donor (category 0) has the highest recall rate (1.00), indicating the possibility of no false negative results. Liver cirrhosis (category 2) and liver fibrosis (category 3) also showed a slight decrease in precision and recall, indicating the possibility of overlapping or ambiguity in these diagnoses. And despite these minor fluctuations, performance in all categories remained unchanged without any significant decline in any of the critical metrics.

The feature importance analysis shows that protein (PROT) is the most influential feature, followed by cholinesterase (CHE), albumin (ALB), and bilirubin (BIL). Unlike the previous dataset, no single feature completely dominates, indicating that many factors play a fundamental role in classification. Age, along with ALT and AST enzymes, plays important roles that contribute to classification, demonstrating their significance in detecting various disorders. This shows that the combination of biochemical indicators and demographic information is essential for accurate classification in this dataset.

On ILPD dataset the model demonstrates excellent performance in predicting liver disease, achieving a high accuracy of 95%, with equally strong precision, recall, and F1-score values. The confusion matrix reveals that both classes are well-classified. Feature importance analysis highlights Total Bilirubin and liver enzymes (e.g., Aspartate and Alkaline Aminotransferase) as the most critical predictors, aligning well with medical understanding of liver function indicators. Overall, the results suggest that the model is both reliable and clinically meaningful, making it a strong candidate for supporting liver disease diagnosis.

4.5 DISCUSSION AND COMPARISON WITH PREVIOUS STUDIES

4.5.1 Studies Used HCV&EGY Datasets

Computer algorithms that aid in discovering trends in the collected information and predictions of different outcomes are based on the data used. As a valuable resource for real-time disease identification and informed medical decision-making, it will serve as a significant tool by integrating AI capabilities to develop predictive models. As can be seen in Table 10, this study's 98% accuracy demonstrates the efficacy of the Extra Trees Classifier combined with SMOTE for HCV prediction.

Table 10: Comparison of the performance across various studies used HCV&EGY datasets.

Study	Dataset used	Aim	Methods	Accuracy
Nandipati 2020 [28]	HCV-EGY Data	HCV Prediction	Preprocessing, feature selection, ML techniques	54.56% (RF)
Syafa'ah et al. 2021 [29]	HCV Data	HCV Prediction	KNN, Naïve Bayes, NN, RF	95.12% (NN)
Ghazal 2021 [30]	HCV-EGY Data	HCV staging Prediction	Fine Gaussian SVM	97.9%
Butt et al. 2021 [32]	HCV-EGY Data	HCV staging Prediction	Artificial Neural Networks (ANN)	98.89%
This Study	HCV Data, HCV-EGY Data	Predict HCV	Extra Trees Classifier with SMOTE	98% (both datasets)

Ghazal (2021) [30] achieved an accuracy of 97.9% with Fine Gaussian SVM, which demonstrates the strength of kernel-based methods in medical data classification. However, SVMs can be computationally expensive and less interpretable compared to ensemble methods like Extra Trees Classifier.

Syafa'ah et al. (2021) [29] achieved 95.12% using Neural Networks, but neural networks require more computational power and hyperparameter tuning, whereas Extra Trees performs well with default settings.

4.5.2 Strengths

- **Importance of Data Preprocessing and Feature Selection:** To guarantee improved generalizability, this study used two datasets (HCV-EGY and HCV), unlike previous research that used a single dataset. Models have been trained Models trained on balanced datasets outperformed those trained on imbalanced datasets. after applying SMOTE to equalize class distributions (e.g., Nandipati et al., 2020) [28]. The feature importance analysis highlighted key biochemical markers (ALT, AST, RBC, RNA levels) that contribute most to classification, enhancing model transparency.

- **Model Selection and Justification:** The Extra Trees Classifier was chosen because of its ability to handle feature redundancy, minimize variance, and increase generalization, making it ideal for biomedical use. Extra Trees Classifier adds more unpredictability in split selection than Random Forest, which frequently results in higher generalization and stability. While ANNs (Butt et al., 2021) [32] got 98.89% accuracy, but to prevent overfitting, they need additional hyperparameter tweaking and bigger datasets.
- **Addressing Overfitting and Generalizability:** Some studies like Nandipati et al. 2020 [28] showed low accuracy (54.56%), mainly due to the imbalanced datasets and insufficient preprocessing. By employing SMOTE for the data balance and the feature importance analysis to avoid reliance on superfluous features, this study reduced the danger of the overfitting. To evaluate the real-world applicability, the future studies should investigate an external validation on the separate datasets.

4.5.3 Studies Used ILPD Dataset

This study can provide a ML for the purpose of predicting hepatic disorders from the ILPD dataset, with performance evaluation against six recent peer-reviewed studies. The proposed method in this project includes advanced data preprocessing techniques, feature selection, and ensemble learning classifiers, resulting in an accuracy of up to 95%. This comparative analysis addresses the strengths and weaknesses of the proposed method in relation to each individual study.

Table 11: Comparison of the performance across various studies used ILPD dataset

Study	Dataset used	Methods	Accuracy
Straw & Wu (2022) [36]	ILPD (UCI)	Sex-stratified analysis of LR, RF, SVM, NB	71.31% (LR), 79.40% (SVM)
Amin et al. (2023) [41]	ILPD (UCI)	PCA + FA + LDA integration, multiple classifiers incl. LR, RF, KNN, SVM, MLP, Voting	88.10%
Md et al. (2023) [42]	ILPD (UCI)	Multivariate imputation, scaling, ensemble classifiers (ETC, RF, etc.)	91.82% (ETC), 86.06% (RF)
Karna et al. (2024) [43]	ILPD (UCI)	LDA, FA, t-SNE, UMAP + classifiers (RF, MLP, KNN, LR)	98.31% (10-fold), 95.79% (train-test)

Table 11 Cont.

Study	Dataset used	Methods	Accuracy
Dashti et al. (2024) [44]	ILPD (UCI)	MLP Neural Networks with backpropagation	99.5%
This study	ILPD (UCI)	SMOTE + ETC with advanced preprocessing and feature selection	95%

Straw and Wu (2022) [36] investigated algorithmic bias in liver disease prediction using ILPD, focusing specifically on sex-stratified performance disparities. Their reported model accuracies ranged from 71.31% (logistic regression) to 79.40% (support vector machine). The relatively lower performance metrics can be attributed to the simplicity of the models used and the study's emphasis on fairness analysis rather than optimization for predictive accuracy. The suggested model, initially engineered for accuracy, furthermore, provides a versatile framework for future fairness auditing, owing to its modular pipeline architecture and availability of feature-level importances.

Amin et al. (2023) [41] presented a model that can be used to extract projection-based statistical features, which may include PCA, Factor Analysis, and then use traditional classifiers for classification. Their methodology achieved an accuracy of 88.10%. While this study demonstrates an excellent strategy for feature integration, it seemingly overlooks addressing the issue of class imbalance or applying advanced sampling strategies and methods. The enhanced performance of this proposed model was due to the integration SMOTE, the application of iterative imputation for missing values, and the selection of flexible ensemble models such as XGBoost and LightGBM, which provide significant superiority.

Md et al. (2023) [42] introduced an enhanced preprocessing strategy employing multiple scaling techniques, multivariate imputation, and ensemble classifiers including Extra Trees, Random Forest, and XGBoost. Their best-performing model, based on the Extra Tree classifier, achieved an accuracy of 91.82%. While methodologically similar to the proposed approach, the slightly higher performance observed in the proposed study can be selective use of PowerTransformer and robust feature selection through cross-validated feature importances, along with a systematic comparison of multiple classifiers under balanced conditions using SMOTE.

Karna et al. (2024) [43] A study was presented with a focus on dimensionality reduction methodologies, including LDA, FA, combined with various classifiers. They reported an accuracy 98.31% by using ten-fold cross-validation 95.79%. Compared to the model's accuracy, these outcomes are marginally better. dimensionality reduction methodology employed, particularly t-SNE and UMAP, are typically used for visualization and are not always reliable for predictive modeling. Their use may introduce instability in model performance depending on parameter settings. In contrast, the present study emphasizes interpretability and stability by using tree-based feature importance and more interpretable ensemble models, which are more appropriate for clinical applications.

Dashti et al. (2024) [44] introduced a self-predictive diagnostic system based on multilayer perceptron (MLP) neural networks, achieving a reported accuracy of 99.5%. While this performance exceeds that of the proposed model, the study lacks sufficient detail on validation protocols, such as cross-validation or external testing. Neural networks, particularly MLPs, are known to perform well on small datasets but are also prone to overfitting if not carefully regularized or validated. In contrast, the present study employs cross-validated feature selection and a hold-out test set, offering a more conservative and generalizable estimate of model performance.

4.6 SUMMARY OF FINDINGS

HCV & EGY datasets exhibit high classification accuracy (98%), indicating strong model performance. The first dataset’s classification is more dependent on a single feature, while the second dataset’s classification is more distributed across multiple features. Misclassifications are minimal in both cases, but the second dataset shows a slightly more distributed misclassification pattern, particularly in Fibrosis and Cirrhosis classifications. Feature importance differs significantly between the two datasets:

The first dataset is strongly influenced by a single feature (Baseline Histological Grading). The second dataset relies on multiple key features (PROT, CHE, ALB, and BIL), meaning the classification decision is based on a more complex feature set.

Class-wise performance is stable in both datasets, but the second dataset experiences minor variations in recall and precision, particularly for Fibrosis and Cirrhosis.

For all of which utilized in the dataset of Indian hepatitis patients. This suggested model achieved an accuracy of 95%, outperforming most existing approaches due to its comprehensive and modern pipeline. This pipeline incorporated advanced preprocessing (iterative imputation, PowerTransformer scaling), feature selection through cross-validated feature importances, class imbalance correction via SMOTE, and evaluation across multiple ensemble classifiers including XGBoost, LightGBM, and Random Forest.

Among the reviewed studies, only two reported higher accuracies: Karna et al. (2024) [43] with 98.31% using a combination of dimensionality reduction techniques, and Dashti et al. (2024) with 99.5% using a multilayer perceptron neural network. However, these methods either involved potentially unstable nonlinear feature transformations or lacked detailed validation, raising concerns about overfitting and generalizability. Other research used older machine learning methods that lacked advancements in data balancing and feature engineering, resulting in 71% to 91% worse performance.

CHAPTER V
CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

This investigation showcases the efficacy of machine learning methodologies in forecasting outcomes related to Hepatitis C Virus (HCV), attaining an impressive 98% accuracy through the application of the Extra Trees Classifier in conjunction with SMOTE. This study employs three unique datasets (HCV-EGY, HCV, and ILPD) and conducts a feature importance analysis to develop the interpretability and ability to be generalized of predictive models for non-invasive HCV diagnosis.

This method outperforms conventional techniques like Random Forest and SVM in HCV prediction, effectively tackles class imbalance, and enhances feature selection insights as compared to previous studies. The results demonstrate how machine learning might help doctors diagnose patients earlier and lessen their need for invasive techniques like liver biopsies.

The deficiencies of the study include the need for external validation on distinct datasets and an increase in field testing of clinical situations. Future research should concentrate on increasing dataset diversity, investigating deep learning methodologies, and incorporating ML-powered decision support systems into healthcare frameworks. This research advances non-invasive diagnostic methods, making HCV detection more accessible and efficient, ultimately facilitating better disease management and treatment planning.

5.2 LIMITATION OF THE STUDY

With the goal of expanding the application of AI algorithms to the prediction of HCV and prognosis, the limitations of the study must be addressed. even if the classification accuracy was high at 98%. heterogeneity in performance between classes, especially in the HCV dataset, where the fibrosis and cirrhosis classifications showed a little decline in accuracy and recall indicates a potential misclassification of disease progression stages, which could impact clinical decision-making and lead to delayed or incorrect interventions.

different levels of reliance on clinical and biochemical markers in different datasets. In the HCV-EGY dataset, the model predominantly depended on a singular dominant feature (Baseline Histological Grading), but the second dataset exhibited a more equitable contribution from other characteristics (PROT, CHE, ALB, BIL, etc.). This inconsistency makes me worry that we might be relying too much on certain biomarkers, which could make the model less useful for diverse groups of patients and in different therapeutic contexts.

supervised learning methods may encounter difficulty in identifying concealed patterns in cases that are under-represented or ambiguous. The model's capacity to accurately represent disease progression over time is further compromised by its dependence on static clinical and biochemical markers rather than longitudinal patient data. Lastly, the model's performance may be compromised by real-world clinical applications, which introduce challenges such as data variability, noise and missing values despite controlled datasets of this proposed method.

Regarding the robust efficacy of this work and model, certain limitations must be considered, the model development and evaluation were based solely on the ILPD dataset, which contains only 583 samples. Small sample size could limit the model generalizability to larger and various populations. Furthermore, the dataset is known to be imbalanced and lacks comprehensive demographic information such as ethnicity, comorbidities, or medication history, which are important factors in clinical diagnosis but were not available for inclusion.

Second, while the model incorporates cross-validation and hold-out testing, it has not yet been externally validated on autonomous datasets and in real-world clinical settings. The absence of extrinsic validation may limit the assessment of its strength across different populations and environments that provide data collection.

Third. Ensemble classifiers like XGBoost and LightGBM can improve accuracy, but they may also reduce interpretability compared to simpler models. In healthcare environments, patients rely on medical staff to explain difficult concepts and make informed decisions. This can pose a problem; while using SMOTE to correct class imbalance, excessive synthetic sampling can lead to the emergence of new artificial patterns that do not accurately represent the actual complexity of patient data. Future research, for efficiency and clinical benefit, they should consider integrating methods for expanding data with biological information more extensively and broadly and adding longitudinal or imaging data.

5.3 FUTURE DIRECTION RESEARCH

Future research should concentrate on enhancing the model robustness to improve health interventions for predicting hepatitis C virus. One of the methods involves exploring ensemble learning techniques, cost-sensitive learning, and new data augmentation strategies to enhance recall in critical categories, specifically liver fibrosis. Additionally, integrating semi-supervised by incorporating supervised or unsupervised learning approaches, because it is more adaptable to various clinical contexts, the model may be better able to identify trends in cases where the answers are ambiguous.

To guarantee model interpretation and generalisability across many datasets, future research should investigate sophisticated methods for feature selection and dimensionality reduction regarding the identification of critical clinical and biochemical markers. Machine learning frameworks should prioritize multi-feature integration to provide a diagnosis that can be more comprehensive progression of HCV infection instead of relying on a single biomarker.

In order to maintain the absence of a gap between experimental accuracy and the real-world application of clinical, future research should study preprocessing techniques, domain adaptation methods, and federated learning methodologies. These advancements will help ensure that machine learning models for HCV prediction are scalable, interpretable, and applicable across different healthcare environments, contributing to better patient outcomes and optimized intervention strategies.

REFERENCES

[1] Wilkins T., MALCOLM J. K., RAINA D., and SCHADE R. R. (2010), "Hepatitis C: Diagnosis and treatment", *American Family Physician*, Vol. 81, No. 11, pp. 1351–1357.

[2] SAFDARI Reza, DEGHATIPOUR Amir, GHOLAMZADEH Marsa, and MAGHOOLI Keivan (2022), "Applying data mining techniques to classify patients with suspected hepatitis C virus infection", *Journal of Intelligent Medicine*, Vol. 2, No. 4, pp. 193–198, , DOI: 10.1016/j.imed.2021.12.003.

[3] AYELDEEN H., SHAKER O., AYELDEEN G., and ANWAR K. M. (2015), "Prediction of liver fibrosis stages by machine learning model: A decision tree approach", *2015 Third World Conference on Complex Systems (WCCS)*, pp. 1-6, Marrakech, Morocco, DOI: 10.1109/ICoCS.2015.7483212.

[4] Suzuki Tetsuro, Ishii Koji, Aizaki Hideki, and Wakita Takaji (2007), “Hepatitis C Viral Life Cycle”, *Journal of Advanced Drug Delivery Reviews*, Vol. 59, No. 12, pp. 1200–1212, DOI: 10.1016/j.addr.2007.04.014.

[5] SOOFI Aized Amin and AWAN Arshad (2017), “Classification Techniques in Machine Learning: Applications and Issues”, *Journal of Basic & Applied Sciences*, Vol. 13, No. 1, pp. 459–465, DOI: 10.6000/1927-5129.2017.13.76.

[6] OSISANWO F.Y., AKINSOLA J.E.T., AWODELE O., HINMIKAIYE J.O., OLAKANMI O., and AKINJOBI J. (2017), “Supervised Machine Learning Algorithms: Classification and Comparison”, *International Journal of Computer Trends and Technology (IJCTT)*, Vol. 48, No. 3, pp. 128–138, DOI: 10.14445/22312803/IJCTT-V48P126.

[7] KOUROU Konstantina, EXARCHOS Themis P., EXARCHOS Konstantinos P., KARAMOUZIS Michalis V., and FOTIADIS Dimitrios I. (2015), “Machine Learning Applications in Cancer Prognosis and Prediction”, *Journal of Computational and Structural Biotechnology*, Vol. 13, pp. 8–17, DOI: 10.1016/j.csbj.2014.11.005.

[8] HABEHH Hafsa and GOHEL Suril (2021), “Machine Learning in Healthcare”, *Journal of Current Genomics*, Vol. 22, No. 4, pp. 291–300, DOI: 10.2174/1389202922666210705124359.

[9] CRUZ Joseph A. and WISHART David S. (2006), “Applications of Machine Learning in Cancer Prediction and Prognosis”, *Journal of Cancer Informatics*, Vol. 2, p. 117693510600200030, DOI: 10.1177/117693510600200030.

[10] HASHEM Somaya, ESMAT Gamal, ELAKEL Wafaa, HABASHY Shahira, ABDEL RAOUF Safaa, ELHEFNAWI Mohamed, ELADAWY Mohamed I., and ELHEFNAWI Mahmoud (2018), “Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients”, *Journal of IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 15, No. 3, pp. 861–868, DOI: 10.1109/TCBB.2017.2690848.

[11] WAKIM Khalil G. (1954), “Physiology of the Liver”, *The American Journal of Medicine*, Vol. 16, No. 2, pp. 256–271, DOI: 10.1016/0002-9343(54)90342-3.

[12] MANNS Michael P., BUTI Maria, GANE Ed, PAWLOTSKY Jean-Michel, RAZAVI Homie, TERRAULT Norah, and YOUNOSSI Zobair (2017), “Hepatitis C Virus Infection”, *Journal of Nature Reviews Disease Primers*, Vol. 3, Article No. 17006, DOI: 10.1038/nrdp.2017.6.

[13] LAUER Georg M. and WALKER Bruce D. (2001), “Hepatitis C Virus Infection”, *New England Journal of Medicine*, Vol. 345, No. 1, pp. 41–52, DOI: 10.1056/NEJM200107053450107.

[14] WESTBROOK Rachel H. and DUSHEIKO Geoffrey (2014), “Natural History of Hepatitis C”, *Journal of Hepatology*, Vol. 61, No 1, pp. S58–S68, DOI: 10.1016/j.jhep.2014.07.012.

[15] CHANDLER Laura (2000), “Diagnostic Tests for Hepatitis C Virus”, *Journal of Clinical Microbiology Newsletter*, Vol. 22, No. 19, pp. 145–149, DOI: 10.1016/S0196-4399(00)80011-2.

[16] NASTESKI Vladimir (2017), “An Overview of the Supervised Machine Learning Methods”, *Journal of HORIZONS.B*, Vol. 4, pp. 51–62, DOI: 10.20544/horizons.b.04.1.17.p05.

[17] JAIN Nipun and KUMAR Rajeev (2022), “A Review on Machine Learning & Its Algorithms”, *International Journal of Soft Computing and Engineering*, Vol. 12, No. 5, pp. 1–5, DOI: 10.35940/ijsce.E3583.1112522.

[18] MOHAMED Amr (2017), “Comparative Study of Four Supervised Machine Learning Techniques for Classification”, *International Journal of Applied Science and Technology*, Vol. 7, No. 2, pp. 5–18.

[19] CHOWDHURY Shovan and SCHOEN Marco P. (2020), “Research Paper Classification using Supervised Machine Learning Techniques”, *In, 2020 Intermountain Engineering, Technology and Computing (IETC)*, pp. 1–6, IEEE, DOI: 10.1109/ietc47856.2020.9249211.

[20] JAVAID Mohd, HALEEM Abid, PRATAP SINGH Ravi, SUMAN Rajiv and RAB Shanay (2022), “Significance of Machine Learning in Healthcare: Features, Pillars and Applications”, *International Journal of Intelligent Networks*, Vol. 3, pp. 58–73, DOI: 10.1016/j.ijin.2022.05.002.

[21] FIELD Kathryn M, DOW Chris and MICHAEL Michael (2008), “Part I: Liver Function in Oncology: Biochemistry and Beyond”, *Journal of The Lancet Oncology*, Vol. 9, No. 11, pp. 1092–1101, DOI: 10.1016/s1470-2045(08)70279-1.

[22] ALTER Miriam J. (2011), “Viral Hepatitis C”, *Journal of Tropical Infectious Diseases: Principles, Pathogens and Practice*, pp. 427–432, DOI: 10.1016/b978-0-7020-3935-5.00065-3.

[23] GHOSH Soumita, ZHAO Xun, ALIM Mouaid, BRUDNO Michael and BHAT Mamatha (2024), “Artificial Intelligence Applied to ‘Omics Data in Liver Disease: Towards a Personalised Approach for Diagnosis, Prognosis and Treatment”, *Gut*, Vol. 74, No. 2, pp. 295–311, DOI: 10.1136/gutjnl-2023-331740.

[24] AHN Joseph C., QURESHI Touseef A., SINGAL Amit G., LI Debiao, and YANG Ju-Dong (2021), “Deep Learning in Hepatocellular Carcinoma: Current Status and Future Perspectives”, *World Journal of Hepatology*, Vol. 13, No. 12, pp. 2039–2051, DOI: 10.4254/wjh.v13.i12.2039.

[25] KONERMAN Monica A., BESTE Lauren A., VAN Tony, LIU Boang, ZHANG Xuefei, ZHU Ji, SAINI Sameer D., SU Grace L., NALLAMOTHU Brahmajee K., IOANNOU George N. and WALJEE Akbar K. (2019), “Machine Learning Models to Predict Disease Progression Among Veterans with Hepatitis C Virus”, Vol. 14, No. 1, pp. e0208141, DOI: 10.1371/journal.pone.0208141.

[26] BARAKAT Nahla H., BARAKAT Sana H. and AHMED Nadia (2019), "Prediction and Staging of Hepatic Fibrosis in Children with Hepatitis C Virus: A Machine Learning Approach", *Healthcare Informatics Research*, Vol. 25, No. 3, pp. 173–181, DOI: 10.4258/hir.2019.25.3.173.

[27] AHAMMED Khair, SATU Md. Shahriare, KHAN Md. Imran and WHAIDUZZAMAN Md. (2020), "Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods", *Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP)*, pp. 1371-1374, Dhaka, Bangladesh, DOI: 10.1109/TENSYMP50017.2020.9230464.

[28] NANDIPATI Sudheer Chandra Rao, XINYING Chen and KHAW Khai Wah (2020), "Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques", *Applications of Modelling and Simulation*, Vol. 4, pp. 89–100.

[29] SYAFA'AH Lailis, ZULFATMAN Zulfatman, PAKAYA Ilham and LESTANDY Merinda (2021), "Comparison of Machine Learning Classification Methods in Hepatitis C Virus", *Jurnal Online Informatika*, Vol. 6, No. 1, pp. 73–78, DOI: 10.15575/join.v6i1.719.

[30] GHAZAL M. Taher, ANAM Marrium, HASAN Mohammad Kamrul, HUSSAIN Muzammil, FAROOQ Muhammad Sajid, ALI Hafiz Muhammad Ammar, AHMAD Munir and SOOMRO Tariq Rahim (2021), "Hep-Pred: Hepatitis C Staging Prediction Using Fine Gaussian SVM", *Computers, Materials & Continua*, Vol. 69, No. 1, pp. 191–203, DOI: 10.32604/cmc.2021.015436.

[31] MOSTAFA Fahad, HASAN Easin, WILLIAMSON Morgan and KHAN Hafiz (2021), "Statistical Machine Learning Approaches to Liver Disease Prediction", *Livers*, Vol. 1, No. 4, pp. 294–312, DOI: 10.3390/livers1040023.

[32] BUTT Muhammad Bilal, ALFAYAD Majed, SAQIB Shazia, KHAN M. A., AHMAD Munir, KHAN Muhammad Adnan and ELMITWALLY Nouh Sabri (2021), "Diagnosing the Stage of Hepatitis C Using Machine Learning", *Journal of Healthcare Engineering*, Vol. 2021, pp. 1–8, DOI: 10.1155/2021/8062410.

[33] PENG Junfeng, ZOU Kaiqiang, ZHOU Mi, TENG Yi, ZHU Xiongyong, ZHANG Feifei and XU Jun (2021), "An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients", *Journal of Medical Systems*, Vol. 45, No. 5, DOI: 10.1007/s10916-021-01736-5.

[34] KAUNANG Fergie (2022), “A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms”, *8ISC Proceedings: Technology,* pp. 33–42, https://ejournal.unklab.ac.id/index.php/8ISCTE/article/view/684, DoA.01.01.2025.

[35] SHINDE Saurabh (2022), “Liver Disease Prediction Based on Grid Search and Random Forest Classification”, *International Journal of Engineering Applied Sciences and Technology*, Vol. 7, No. 1, pp. 136–140, DOI: 10.33564/ijeast.2022.v07i01.020.

[36] STRAW Isabel and WU Honghan (2022), “Investigating for Bias in Healthcare Algorithms: A Sex-Stratified Analysis of Supervised Machine Learning Models in Liver Disease Prediction”, *BMJ Health & Care Informatics*, Vol. 29, No. 1, pp. e100457, DOI: 10.1136/bmjhci-2021-100457.

[37] SACHDEVA Ravi Kumar, BATHLA Priyanka, RANI Pooja, SOLANKI Vikas and AHUJA Rakesh (2023), “A Systematic Method for Diagnosis of Hepatitis Disease Using Machine Learning”, *Innovations in Systems and Software Engineering*, Vol. 19, No. 1, pp. 71-80, DOI: 10.1007/s11334-022-00509-8.

[38] KIM Sun–Hwa, PARK So–Hyeon and LEE Heeyoung (2023), “Machine Learning for Predicting Hepatitis B or C Virus Infection in Diabetic Patients”, *Scientific Reports*, Vol. 13, No. 1, DOI: 10.1038/s41598-023-49046-9.

[39] ALOTAIBI Abrar, ALNAJRANI Lujain, ALSHEIKH Nawal, ALANAZY Alhatoon, ALSHAMMASI Salam, ALMUSAIRII Meshael, ALRASSAN Shoog and ALANSARI Aisha (2023), “Explainable Ensemble-Based Machine Learning Models for Detecting the Presence of Cirrhosis in Hepatitis C Patients”, *Computation*, Vol. 11, No. 6, pp. 104, DOI: 10.3390/computation11060104.

[40] ALI Ali Mohd, HASSAN Mohammad R., ABURUB Faisal, ALAUTHMAN Mohammad, ALDWEESH Amjad, AL-QEREM Ahmad, JEBREEN Issam and NABOT Ahmad (2023), “Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection”, *Machines*, Vol. 11, No. 3, pp. 391, DOI: 10.3390/machines11030391.

[41] AMIN Ruhul, YASMIN Rubia, RUHI Sabba, RAHMAN Md Habibur and REZA Md Shamim (2023), "Prediction of Chronic Liver Disease Patients Using Integrated Projection-Based Statistical Feature Extraction with Machine Learning Algorithms", *Informatics in Medicine Unlocked*, Vol. 36, Article 101155, DOI: 10.1016/j.imu.2022.101155.

[42] MD Abdul Quadir, KULKARNI Sanika, JOSHUA Christy Jackson, VAICHOLE Tejas, MOHAN Senthilkumar and IWENDI Celestine (2023), "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease", *Biomedicines*, Vol. 11, No. 2, Article 581, DOI: 10.3390/biomedicines11020581.

[43] KARNA Anand, KHAN Naina, RAUNIYAR Rahul and SHAMBHARKAR Prashant Giridhar (2024), "Unified Dimensionality Reduction Techniques in Chronic Liver Disease Detection", *arXiv preprint,* DOI: 10.48550/arXiv.2412.21156.

[44] DASHTI Fatemeh, GHAFFARI Ali, SEYFOLLAHI Ali and ARASTEH Bahman (2024), "A Self-Predictive Diagnosis System of Liver Failure Based on Multilayer Neural Networks", *Multimedia Tools and Applications*, Vol. 83, No. 36, pp. 83769–83788, DOI: 10.1007/s11042-024-18945-y.

[45] JIANG Tammy, GRADUS Jaimie Lauren and ROSELLINI Anthony James (2020), "Supervised Machine Learning: A Brief Primer", *Behavior Therapy*, Vol. 51, No. 5, pp. 675–687, DOI: 10.1016/j.beth.2020.05.002.

[46] CHOUDHARY Rishabh and GIANEY Hemant Kumar (2017), "Comprehensive Review on Supervised Machine Learning Algorithms", *In*, *Proceedings of the 2017 International Conference on Machine Learning and Data Science (MLDS)*, pp. 37–43, IEEE, DOI: 10.1109/MLDS.2017.11.

[47] RAHMAN Mohammad Mostafizur and DAVIS Derek Norman (2013), "Addressing the Class Imbalance Problem in Medical Datasets", *International Journal of Machine Learning and Computing*, Vol. 3, No. 3, pp. 224–228, DOI: 10.7763/ijmlc.2013.v3.307.

[48] PERANGINANGIN Resianta, HARIANJA Eva Julia Gunawati, JAYA Indra Kelana and RUMAHORBO Benget (2020), “PENERAPAN ALGORITMA SAFE-LEVEL-SMOTE UNTUK PENINGKATAN NILAI G-MEAN DALAM KLASIFIKASI DATA TIDAK SEIMBANG”, *METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi*, Vol. 4, No. 1, pp. 67–72, DOI: 10.46880/jmika.vol4no1.pp67-72.

[49] SEO Jae-Hyun and KIM Yong-Hyuk (2018), “Machine-Learning Approach to Optimize SMOTE Ratio in Class Imbalance Dataset for Intrusion Detection”, *Computational Intelligence and Neuroscience*, Vol. 2018, pp. 1–11, DOI: 10.1155/2018/9704672.

[50] TU My Chau, SHIN Dongil and SHIN Dongkyoo (2009), “A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms”, *In, Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC)*, pp. 183–187, IEEE, DOI: 10.1109/dasc.2009.40.

[51] NGUYEN Thanh, KHOSRAVI Abbas, CREIGHTON Douglas and NAHAVANDI Saeid (2015), “Medical data classification using interval type-2 fuzzy logic system and wavelets”, *Applied Soft Computing*, Vol. 30, pp. 812–822, DOI: 10.1016/j.asoc.2015.02.016.

[52] ZHANG Lin, WANG Jixin, CHANG Rui and WANG Weigang (2024), “Investigation of the effectiveness of a classification method based on improved DAE feature extraction for hepatitis C prediction”, *Scientific Reports*, Vol. 14, No. 1, Article 6457, DOI: 10.1038/s41598-024-59785-y.

[53] VENKATESH Bhuvaneswari and ANURADHA Jeyalakshmi (2019), “A Review of Feature Selection and Its Methods”, *Cybernetics and Information Technologies*, Vol. 19, No. 1, pp. 3-26, DOI: 10.2478/cait-2019-0001.

[54] KUMAR Vipin and MINZ Rakesh (2014), “Feature Selection: A Literature Review”, *The Smart Computing Review*, Vol. 4, No. 3, pp. 211–229, DOI: 10.6029/smartcr.2014.03.007.

[55] JOVIC Aleksandar, BRKIC Kresimir and BOGUNOVIC Nikola (2015), “A Review of Feature Selection Methods with Applications”, *In, Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, IEEE, DOI: 10.1109/MIPRO.2015.7160458.

[56] CHANDRASHEKAR Girish and SAHIN Ferat (2014), “A survey on feature selection methods”, *Computers and Electrical Engineering*, Vol. 40, No. 1, pp. 16–28, DOI: 10.1016/j.compeleceng.2013.11.024.

[57] AGGARWAL Charu C. (2015), *Data Classification*: *Algorithms and Applications*, First Edition, CRC Press (Taylor & Francis Group), Boca Raton, Florida.

[58] KOTSIANTIS Sotiris Billas, ZAHARAKIS Ioannis Dimitrios and PINTELAS Panagiotis Evangelos (2006), “Machine learning: A review of classification and combining techniques”, *Artificial Intelligence Review*, Vol. 26, No. 3, pp. 159–190, DOI: 10.1007/s10462-007-9052-3.

[59] CORMACK Robert Melville (1971), “A Review of Classification”, *Journal of the Royal Statistical Society. Series A (General)*, Vol. 134, No. 3, pp. 321, DOI: 10.2307/2344237.

[60] GEURTS Pierre, ERNST Damien and WEHENKEL Louis (2006), “Extremely randomized trees”, *Machine Learning*, Vol. 63, No. 1, pp. 3–42, DOI: 10.1007/s10994-006-6226-1.

[61] YAĞANOĞLU Mete and KÖSE Cemal (2018), “Real-Time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People”, *Electronics*, Vol. 7, No. 4, pp. 50, DOI: 10.3390/electronics7040050.

[62] POWERS David Martin William (2020), “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*”, arXiv preprint,* DOI: 10.48550/arXiv.2010.16061.

[63] ÇELEN Mustafa K., ERTÜRK ŞENGEL Buket, KAYA Şafak, DEMİRTÜRK Neşe, AZAP Alpay, PULLUKÇU Hüsnü, EROĞLU Esma, YILDIRIM Figen, BARUT Hüseyin Ş., ZERDALI Esra, SAĞMAK TARTAR Ayşe, METE Ayşe Ö., ŞAHİN Ahmet M., MUTAY SUNTUR Bedia, SARI Nagehan D., YILMAZ Emel, CANDEVİR Aslıhan, ŞİMŞEK Funda, İNAN Dilara, AKHAN Sıla, ASAN Ali, GÜNAL Özgür, URAL Onur, PARLAK Mehmet, ÇABALAK Mehmet, NAZİK Selçuk, HIZEL Kenan, KINIKLI Sami, BEŞTEPE DURSUN Zehra, BATIREL Ayşe and MERMUTLUOĞLU Çiğdem (2024), "Treatment Initiation Rates of Patients with Positive Anti-Hepatitis C Virus Results in Tertiary Hospitals in Turkey", *The Journal of Infection in Developing Countries*, Vol. 18, No. 3, pp. 441–449DOI: 10.3855/jidc.17910.

[64] KHUDHAIR Hasan Abd Ali, ALBAKAA Ali Abdul Hadi and HUSSEIN Khwam Rasheed (2023), "Detecting the Prevalence of Hepatitis C Virus among Iraqi People", *International Journal of Biomedicine*, Vol. 13, No. 2, pp. 234–240, DOI: 10.21103/article13(2)_oa5.

[65] TRICKEY Adam, ARTENIE Adelina, FELD Jordan J. and VICKERMAN Peter (2025), "Estimating the annual number of hepatitis C virus infections through vertical transmission at country, regional, and global levels: a data synthesis study", *The Lancet Gastroenterology and Hepatology*, Vol. 10, No. 7, pp. 551–562, DOI: 10.1016/S2468-1253(25)00189-X.

[66] KAMAL Sanaa, ELELEIMY Mohamed, HEGAZY Doaa and NASR Mahmoud (2017), "Hepatitis C Virus (HCV) for Egyptian Patients", *UCI Machine Learning Repository*, https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients, DoA. 25.01.2024, DOI:10.24432/C5989V.

[67] LICHTINGHAGEN Ralf, KLAWONN Frank and HOFFMANN Georg (2020), "HCV Data", *UCI Machine Learning Repository*, https://archive.ics.uci.edu/dataset/571/hcv+data, DoA. 15.05.2024, DOI:10.24432/C5D612.

[68] RAMANA Bendi and VENKATESWARLU N. (2022), "ILPD (Indian Liver Patient Dataset)", *UCI Machine Learning Repository*, https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset, DoA. 27.08.2024, DOI:10.24432/C5D02C.