A Machine Learning Framework for Hepatic Health Analysis

By
Md. Ashraful Islam Albi
ID: 212-15-4117

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Narayan Ranjan Chakraborty
Associate Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by

Ms. Dristi Saha
Lecturer
Department of Computer Science and Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 17, 2025

APPROVAL

This Project titled **"A Machine Learning Framework for Hepatic Health Analysis,"** submitted by **Md. Ashraful Islam Albi, ID No: 212-15-4117** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **17 September 2025**.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque — Board Chairman
Professor & Associate Head
Department of CSE, FSIT
Daffodil International University

Ms. Nazmun Nessa Moon — Internal Examiner 1
Associate Professor
Department of CSE, FSIT
Daffodil International University

Md. Abbas Ali Khan — Internal Examiner 2
Assistant Professor
Department of CSE, FSIT
Daffodil International University

Dr. Md. Zulfiker Mahmud — External Examiner
Professor
Department of CSE
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of Narayan Ranjan Chakraborty, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

for

Narayan Ranjan Chakraborty
Associate Professor
Department of Computer Science and
Engineering Daffodil International University

Co-Supervised by:

Ms. Dristi Saha
Lecture
Department of Computer Science and
Engineering Daffodil International University

Submitted by:

Ashraful

Md. Ashraful Islam Albi
212-15-4117
Department of Computer Science and
Engineering Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. I'm deeply grateful to everyone who has assisted us in one way or another.

First, I express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for me to complete the **Final Year Design Project (FYDP)** successfully.

I'm grateful and wish my profound indebtedness to **Narayan Ranjan Chakraborty, Associate Professor,** Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of my supervisor in the field of **Machine learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank my entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Liver diseases such as Hepatitis, Cirrhosis, Fatty Liver Disease (NAFLD), and Cholestasis are the main causes of world-wide health burden due to their non-symptomatic presentation and diagnosis at late stages. Correct diagnosis at an earlier stage is crucial to have improved patient outcomes. The present study presents a Machine Learning Framework to characterize the liver health by using supervised methods to categorize and sub-type liver pathology. The data is based on Kaggle's Liver Patient Dataset and includes clinical and biochemical parameters such as Age, Gender, levels of Bilirubin, Alkaline Phosphatase, AST, SGPT, levels of Albumin, and the A/G Ratio. A standard data pre-processing routine was used to clean up the missing values, remove outliers, add new result column and normalize features to prepare and transform the data to feed into models. Its rule-based labeling system identified the categories of the diseases and then several machine learning classifiers were trained on the dataset by employing Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and an ensemble Voting Classifier. The RF and the KNN demonstrated extremely high accuracy (65-99%), but the Voting Classifier demonstrated higher robustness by using ensemble learning. This rule-based system was employed to further classify Hepatitis cases into subtypes: Acute Viral (A/E), Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, and NASH (Fatty Hepatitis). The hybrid system has been implemented as an online interactable web app through Streamlit with real-time liver disease and Hepatitis subtype predictions so it is readily accessible to clinicians, researchers, and patients. This system offers faster early detection and a better liver health diagnostic tool.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Background to the increasing world incidence of liver diseases and the need for advanced techniques for early diagnosis. It clarifies the motivation, objectives, and methodology for using machine learning to enhance early diagnosis and demonstrates the proposed project outcomes. The chapter also provides an overview of the report structure, which acts as the introduction to the rest of the sections.

1.1 Introduction

Liver disease has become a rising global healthcare challenge, having affected millions of people worldwide every year and remaining among the leading causes of morbidity and mortality. As the largest internal organ of the body, the liver contains key metabolic processing functions, detoxification, protein synthesis, and immunity. Functions of the liver, such as Hepatitis, Cirrhosis, Fatty Liver Disease, and Cholestasis critically incapacitate these functions and create life-risking conditions if not diagnosed at an early stage. The problem is that liver disease is normally asymptomatic at early phases of its onset, consequently, forestalling diagnosis and treatment. Thus, global healthcare systems face increasing pressure to devise more accurate, efficient, and more accessible diagnosis regimens.

Traditional diagnosis approaches like blood work, imaging, and liver biopsies, although effective, are often slow, expensive, or invasive. Also, test interpretations from clinical tests usually require technical knowledge, not readily available in resource-challenged regions. Since AI and data-based methods have been rising exponentially, machine learning has been described as an excellent resource to aid medical diagnosis by uncovering hidden patterns from complex collections of information. The methodologies have the potential of bridging clinical expertise gaps, reducing diagnostic delays, and providing patient-specific information.

The paper, “A Machine Learning Framework for Hepatic Health Analysis”, transcends these challenges by applying machine learning models to classify and predict liver conditions based on clinical and biochemical features. The characteristic set applied for the purpose includes noteworthy characteristics such as Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, AST (SGOT), ALT (SGPT), Total Proteins,

Albumin, and A/G Ratio. With the use of preprocessing, feature construction, and rule-based classification, the paper develops warranted disease classes such as Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain. Furthermore, the framework continues to advance the analysis by dividing Hepatitis into finer levels such as Acute Viral Hepatitis, Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, and NASH. For robustness purposes, we attempted supervised ML algorithms like Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) and ensemble Voting Classifier. Trained classifiers had significant predicting ability and Voting Classifiers presented the highest reliability. Besides, executing the system in the form of a web application using Streamlit makes the system easily accessible to clinicians and researchers and offers real-time predicting of the disease using an interactive user system. The paper presents the possibility to improve hepatic healthcare by applying machine learning to bridge the gap from traditional diagnosis towards intelligent decision-support systems and to pave the way towards liver disease diagnosis in the early stages at low expenditure and high reliability

1.2 Motivation

Liver problems are a global health problem and have been accountable for millions of deaths every year. The diseases like Hepatitis, Cirrhosis, Fatty Liver (NAFLD), and Cholestasis have worldwide prevalence but diagnosis of these diseases in most developing countries like Bangladesh strongly relies on limited resources. The traditional diagnostic approaches like imaging modalities and liver biopsy are expensive, invasive in nature, and sometimes unavailable. The conventional biochemical blood analysis necessitates sophisticated interpretation by highly skilled hands, prone to human error and interindividual variability. In recent times, machine learning (ML) has been a paradigm-changer in clinical medicine, offering automated, scalable, and reliable diagnostic platforms. As compared to rule-based systems, ML-based systems possess an advantage of deriving the underlying patterns of clinical information, serum bilirubin, enzyme function (AST, ALT, ALP), and protein ratios that might present prefailure symptoms of liver damage. For such applications, ML not only allows physicians to deliver rapid and reliable decisions but is equally beneficial in allowing mass screening campaigns on an unparalleled massive scale in deprived conditions of resources.

1.3 Objectives

The broad objective of the current investigation is to design and develop a machine learning system to diagnose hepatic illnesses in such a way so as to be efficient in automated identification of multiple types of liver pathology but to be beneficial in clinical decision-making too. The special objective of the current effort is to specifically design a system that is not only capable of discriminating Hepatitis, Cirrhosis, Fatty Liver (NAFLD), Cholestasis and distinguishing healthy cases on the basis of common biochemical and demographical parameters but is capable of sub-classification of Hepatitis into dominant types., Acute Viral Hepatitis, Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, and NASH with a purpose to extract deeper information relevant to individualized plans of therapy. The work sets its target on processing the data for pre-conditioning, feature engineering, and cleaning of missing and outlier samples such that it ends up with a clean and reliable dataset useful for model building. Several machine learning classifiers like Random Forest, K-Nearest Neighbors, Support Vector Classifier, and an ensemble Voting Classifier have been implemented, compared, and optimized and tried for quantifying their efficiency in the parameters of accuracy, precision, recall, and F1-score. Additionally, the best suited model is wrapped as an interactive web-based program using Streamlit such that it easily reaches healthcare professionals and patients wherever limited access exists for its use. Finally, enabling interpretability and clinical usefulness of the system is provided special attention such that predictions not merely prove to be correct but reliable and thereby fill the void between high-end computation-based methodologies and real-world healthcare application.

1.4 Methodology

The approach used here is customized for the construction of a framework for hepatic disease diagnosis using machine learning that effectively discriminates liver disease and subtypes of hepatitis from clinical and biochemical features. The approach consists of using a combination of rigorous preprocessing of the used data, feature engineering, applying various machine learning models, performance analysis, and deployment of the framework as a clinical decision support. The initial step involves gathering and preprocessing the data whereby the dataset is scanned for missing value, outliers, and inconsistencies. Non-numerical entities such as “N/A” or “?” are replaced by null and

appropriate imputation or dropping of such observations is carried out for making the information reliable. Redundant attributes such as irrelevant columns are dropped and relevant biochemical attributes such as Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase (ALP), SGOT, SGPT, Total Proteins, Albumin, and A/G Ratio are retained along with demography-based attributes such as Age and Gender. Data normalizing and encoding such as converting Gender to numeric type by mapping Gender to numeric type is carried out for making the dataset eligible for applying machine learning algorithms. After that, a system of disease classification is used with rules that are clinically well-founded such that they will be able to classify cases into classes of Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain. For those patients having Hepatitis, a sub-classification function is incorporated for inferring specific types like Acute Viral Hepatitis (A/E), Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, and NASH. The framework is applied clinically through an application of a two-level system of classifying the disease. The dataset is split into the training and test sets through stratified sampling so as to preserve the class balance. A few machine learning classifiers like Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) and ensemble Voting Classifier have been trained. Hyperparameter optimization and cross-validation have also been used for optimization of performance and avoidance of overfitting. The performance of the model is studied on the basis of measures of classification such as accuracy, precision, recall, F1-score and confusion matrices for evaluation of overall prediction potential. Finally, the best-performing model is presented as an interactive web application using Streamlit such that the end users—clinicians, researchers, or patients can input biochemical test results and get real-time diagnostic predictions. The application also provides information on the input guidelines for convenience. The system ensures predictions are clinically valid, interpretable, and transferable for real-world healthcare practice, in particular resource-limited settings.

1.5 Project Outcome

The findings of these studies confirm the successful development of a machine learning diagnostic system capable of processing biochemical and demographical data and identifying liver disease with more than 96% accuracy. Pitted against and using a variety of different algorithms Random Forest, K-Nearest Neighbors, Support Vector Classifier,

and Voting Classifier combination, a system averaged 96% to 98% overall accuracy and showed good reliability in prediction. Another of its significant contributions is that it not only identifies general liver disease such as Hepatitis, Cirrhosis, Fatty Liver (NAFLD), and Cholestasis but also sub-classifies Hepatitis into six classes (Acute Viral, Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, and NASH). The degree of detail provided by such models is clinically more useful than conventional binary or multiclass models that only give predictions of disease presence versus absence. In addition to accuracy, the outcome of the project comprises the release of the model as an interactive Streamlit web application, so that clinicians and patients may provide biochemical test results and receive predictions on the spot. The application includes rules for obligatory inputs, boosting accessibility and ease of use. By including interpretability modules and short-form classification outputs, the system is designed to aid clinical decision-making in both advanced healthcare institutions and resource-challenged sites. Finally, the project provides an efficient, accessible, and scalable solution that is able to facilitate early identification, ease reliance on invasive diagnosis, and provide healthcare professionals with actionable insights, bridging the gap between computational intelligence and real-world healthcare delivery.

1.6 Organization of the Report

The project consists of six chapters. Chapter 1 provides an introduction to the project, including background, motivation, and objectives and ends with a summary of methodology and expected project results. Chapter 2 provides background in the form of an introduction and a proper literature survey, analyzing literature works, related research work, similar applications, and conducting a gap analysis to outline the research problem. Chapter 3 describes the research methodology, including the proposed framework, functional and non-functional requirements, context diagram, data flow diagram (at Level 1), user interface design, detailed methodology, system design, project plan, and task allocation. Chapter 4 provides environment setup, implementation methodology, testing approach, and performance analysis, demonstrating experimental results due to numerous machine learning models and comparison among them. Chapter 5 states engineering requirements and design challenges encountered by developers, like software, hardware, and communication requirements. It also includes societal and environmental impacts, ethical issues, sustainability issues, project management

practices, and financial analysis. Chapter 6 finally summarizes the project with detailed results summary, limitations of the work, and possible future avenues for research.

Chapter 2

Background

It reviews the past studies and technologies and observes the challenges for the liver disease diagnosis and the hope of machine learning overcoming such challenges. The chapter provides the background for the suggested solution and uncovers gaps in previous.

2.1 Introduction

Liver disorders are among the most serious global health issues and cause millions of global mortalities annually. The liver has fundamental functions such as detoxification, synthesis of proteins and production of bile, and its failure causes serious complications. Common disorders include Hepatitis, Cirrhosis, Fatty Liver Disease (NAFLD), and Cholestasis, many of which remain undetected until late stages. Early diagnosis is, therefore, paramount for reducing mortality and improving patient survival. Conventional diagnosis by use of biopsies and imaging, whilst efficient, is prohibitive and intrusive. Blood tests for bilirubin level, liver enzyme level (SGOT, SGPT, ALP), proteins, and A/G ratio provide valuable information but require expert interpretation, which is not available for resource-deprived localities. Machine learning (ML) offers an efficient solution by taking in medical data, uncovering hidden patterns, and predicting disease classes efficiently. Compared to conventional rule-based systems, ML models automatically optimize and enhance performance by learning from experience and, for that matter, perform exceptionally well for clinical decision-making. This project proposes a Machine Learning Framework for Hepatic Health Analysis that is capable of categorizing overriding liver disorders and further sub-classify Hepatitis into six types of Acute Viral, Chronic B, Hepatitis C, Alcoholic, Autoimmune, and NASH. By using random forest algorithm, KNN, SVC, and Voting Classifiers among others, the framework yields robust predictions. Finally, the system is encapsulated as a Streamlit-based web application and conveniently used and accessed from advanced as well as resource-deprived healthcare settings.

2.2 Literature Review

Table 1: Comparative Analysis of earlier Research Work

Authors	Year	Title	Methodology	Key Findings
Sarvestany et al.	2022	Ensemble ML for Advanced Fibrosis	Ensemble models (SVM, RF, GBM, LR, NN) validated with 1,703 biopsies	AUROC 0.87; outperformed APRI, FIB-4, and NFS; close to elastography
Schattenberg et al.	2021	AI in Hepatology	Comprehensive review of AI in imaging, histopathology, biomarkers	Highlighted automation, validation, and regulation challenges
Yang et al.	2023	NAFLD Risk Prediction with Tabular Data	Seven ML models trained on 21 clinical features	RF, LightGBM, XGBoost AUC ≥0.95 (train), ~0.90 (test)
Ganie et al.	2024	Improved Liver Disease Prediction	Ensemble + feature selection (LASSO) on ILPD	Feature selection improved model stability and performance
UCI Repository	—	Indian Liver Patient Dataset	Dataset with 583 samples and 10 clinical features	Standard benchmark for liver disease ML models
Cheng et al.	2023	Handling Imbalanced Liver Datasets	SMOTE + ENN resampling with RF and KNN	Improved recall and F1 for minority classes
Ma et al.	2025	Hybrid Ensemble for Liver and Heart Disease	Hybrid ensemble (RF, SVM, XGB)	Hybrid achieved >95% accuracy
Ganie et al.	2024	Gradient Boosting for Cirrhosis	Compared XGBoost, LightGBM, CatBoost	Gradient boosting yielded higher AUC
Lim et al.	2024	Clinical Decision Support for NAFLD	EHR longitudinal lab data, external validation	Proved clinical utility of lab-based ML predictors
Spann et al.	2020	Explainability in Hepatic AI	Discussed SHAP and	Highlighted need for

			LIME interpretability	explainability in healthcare
Ratziu et al.	2024	Digital Pathology for NASH	AI-enhanced pathology, cross-validation, external validation	Stressed robust validation practices
Sarvestany et al.	2022	Fibrosis Prediction with Labs	Labs vs elastography for staging	Lab-based ML achieved AUROC >0.85
Ahosan et al.	2025	HBV Outcome Prediction	Ensemble ML for hepatitis B outcomes	Accuracy >90%
Amin et al.	2023	Feature Selection Impact	LASSO + RF applied to ILPD	Reduced overfitting, improved model stability
Rehman et al.	2024	Multi-disease ML Framework	End-to-end cleaning + ensemble pipeline	Worked across multiple datasets
Singh et al.	2022	Deep Learning in Hepatology	Overview of CNNs, RNNs applied to hepatology	Promising results with lab data
Shousha et al.	2023	Risk Stratification in Chronic Hepatitis	RF vs Logistic Regression	RF better at stratifying disease progression risk
Yip et al.	2021	Integration of Biopsy and ML	Combined biopsy + labs	Improved cirrhosis and fibrosis diagnosis
Lou et al.	2020	Explainable Boosting for Liver	Explainable Boosting Machine (EBM) on ILPD	High performance + interpretability
Asrani et al.	2022	AI in Global Liver Disease Burden	AI and epidemiology review	Advocated ML for low-resource screening using lab data

2.2.1 Similar Applications

Sarvestany et al. [1] Ensemble ML for advanced fibrosis (clinical labs + demographics) constructed and validated an ensemble of machine-learning models (SVM, RF, GBM, LR, NN + ensemble) for predicting all-cause advanced hepatic fibrosis using routinely accessible labs and clinical information from 1,703 biopsies and included external validations. The ensemble made 100% determinate classifications and AUROC of 0.87 at most on an inner set and considerably surpassed APRI, FIB-4, and NFS, and came close to transient elastography and expert panels. This paper is a cornerstone showing that tabular, non-imaging clinical features have the promise of powering liver risk models that apply directly to your lab-based setup.

Schattenberg et al. [2] State-of-the-art review: AI implementations throughout imaging, histopathology, non-invasive biomarkers, and prediction models for hepatology. Pros and cons (automation, risk stratification vs. external validation, bias, regulation, clinical uptake) justify multiple algorithm use: thoughtful validation, a path to clinical usability (my Streamlit decision-support app).

Yang et al. [3] NAFLD risk prediction with tabular data educated and contrasted seven ML models on 21 clinical characteristics to forecast NAFLD. RF, LightGBM, and XGBoost achieved AUC ≥ 0.95 on training and ~0.90 on an internal test set, and they highlight that normal labs by themselves have excellent discriminative capability without imaging. Their comparison-of-models technique justifies and fits well with your methodology of thinking of considering multiple learners and an ensemble.

Ganie et al. [4] Improved liver disease prediction from clinical data (BMC Med Inform Decis Mak, 2024) reported ensemble learning using feature selection (eg, LASSO) greatly improves liver set classification (including ILPD). The authors provide optimal algorithm performance changes dependent on preprocessing choices (eg, DT performed best after LASSO) and emphasize the attention in your pipeline on cleaning, feature engineering, and model selection.

UCI Repository [5] ILPD (Indian Liver Patient Dataset) as a benchmark UCI ILPD dataset (583 instances; Age, Gender, TBil, DBil, ALP, SGOT, SGPT, Total Proteins, Albumin, A/G ratio) remains the baseline tabular benchmark for liver disease prediction. The dataset's feature schema is analogous to inputs of your project and is amenableto

comparison of methods, ablation analysis, and external references for your thesis.

Cheng et al. [6] Handling Imbalanced Liver Datasets studied skewed class distributions of ILPD and showed that SMOTE + ENN resampling and RF and KNNimprove F1 and recall of minority classes (e.g., healthy). This validates your decision to preprocess data before training.

Ma et al. [7] Hybrid/Ensemble Frameworks suggesteda hybrid ensemble of RF, SVM, and XGB for liver and heart disease prediction at accuracies of more than 95%. The authors state that hybrids outperform solitarymodels, which lends support for your VotingClassifier solution.

Ganie et al. [8] Gradient Boosting for Cirrhosis compared XGBoost, LightGBM, and CatBoost for the prediction of chronic liver disease. Gradient boosting always produced improved AUC compared to linear/SVM models and checking ensemble tree importance.

Lim et al. [9] Clinical Decision Support for NAFLD externally validated ML models for predicting onset of NAFLD from EHR lab longitudinal data. Their external validation established clinical usefulness of lab-based ML predictors an extremely important consideration for your real-world application.

Spann et al. [10] Explainability in Hepatic AI described the need for explainable AI in hepatology. Model-agnostic explainers of laboratory-based predictions (SHAP/LIME) have been proposed, so that doctors know why a patient is predicted to be hepatitis or cirrhosis.

Ratziu et al. [11] Digital Pathology & Cross-Validation provided an overview of AI-enhanced digital pathology for NASH and described cross-validation and external validation in the service of generalizability. This provides rationale for applying stratified train/test split and CV.
Sarvestany et al. [12] Fibrosis Prediction with Labs Sarvestany showed labs + ML compete against imaging (elastography) for fibrosis staging. Their ensemble achieved AUROC >0.85, showing that for most use cases, non-invasive blood tests are enough.

Ahosan et al. [13] HBV Outcome Prediction Ensemble learning was applied for predicting hepatitis B outcome with >90% accuracy. This is ml-based, their work confirms the use of

ensemble ml for hepatitis-specific results.

Amin et al. [14] Feature Selection Impact utilized LASSO + RF on ILPD and demonstrated that performance is improved and overfitting is decreased by feature selection. This supports your process of eliminating extreme values (TBil / DBil).

Rehman et al. [15] Multi-disease Frameworks built MaLLiDD as a multi-disease ML pipeline for liver disease and cross-validated on multiple sets. Their technique follows that of your end-to-end architecture (cleaning then feature engineering then ensemble of ensembles).

Singh et al. [16] Deep Learning in Hepatology provided an overview of use of deep learning models (CNNs, RNNs) for hepatology. While imaging was predominant, they also indicated promise in systematic laboratory data, and they indicated possibilities for combination of hybrid deep models with table-based inputs of biochemical parameters.

Shousha et al. [17] Risk Stratification in Chronic Hepatitis developed ML models to stratify hepatitis B and C patients into groups of progression risk using common blood markers. Their random forest model provided better prediction of risk compared to logistic regression in facilitating the role of progression risk in ML frameworks.

Yip et al. [18] Integration of Biopsy and ML Model demonstrated that by combining biopsy histology features and clinical data in ML models, they have improved performance for fibrosis and cirrhosis diagnosis. The paper provides an illustration of the readiness of ML frameworks to integrate an assortment of modalities.

Lou et al. [19] Explainable Boosting Models (EBM) for Liver Prediction applied interpretable ML (Explainable Boosting Machine) to ILPD and achieved near-black-box model performance without compromising transparency. This gives credence to interpretable ML application in healthcare, pursuant to the ethical dimension of your project.

Asrani et al. [20] AI in Global Liver Disease Burden linked global liver disease burden to the need for AI-based screening platforms. They argue that ML can have application in resource-constrained areas based on non-invasive laboratory data, exactly similar to how

you propose using such a framework in Bangladesh and other places.

2.3 Gap Analysis

This work contributes to the field by transcending theoretical investigations to provide an exhaustive, clinically oriented, and user-friendly diagnostic system, filling the gap that separates scholarly research from healthcare practice.

Table 2: Gap Analysis of Research vs. Proposed System

Research Feature	Sarvestany et al.	Yang et al.	Ganie et al.	Lim et al.	Spann et al.	Rehman et al.	Proposed System
Ensemble ML for Hepatic Prediction	Yes	No	Yes	No	No	Yes	Yes
Detects Multiple Liver Diseases	No	No	No	Yes	No	Yes	Yes
Hepatitis Sub-type Classification	No	No	No	No	No	No	Yes
Handles Imbalanced Data (SMOTE/ENN)	No	No	Yes	No	No	No	Yes
Explainability (SHAP/LIME/EBM)	No	No	No	No	Yes	No	Yes
External Validation	Yes	No	No	Yes	Yes	No	Yes
Web-based Clinical Decision Support App	No	No	No	No	No	No	Yes

2.4 Summary

The literature survey stresses that a vast quantity of work has been carried out by applying machine learning methodologies for the diagnosis of liver disease. The

preceding works have shown that models such as Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machines have been applied to achieve high precision by applying clinical and biochemical parameters such as bilirubin, liver enzymes such as SGOT, SGPT, ALP, and albumin and protein ratios. Almost all of these works have limited scope and have contemplated a mere single or double liver disorder, without providing a holistic diagnostic solution. One notable commentary from the literature is that Hepatitis is often thought of as a single category in earlier work, despite the fact that it contains numerous subtypes which contain different clinical interventions. Similarly, many earlier works work on small, imbalanced sets, and they have issues of overfitting and high accuracy measures that depress corresponding reliability for real-world use. Moreover, such models have desirable outcomes in laboratory experiments but limited translation to real-world practical tools or user-facing interfaces that clinicians and patients may utilize. The Gap Analysis revealed the following deficiencies: (1) absence of multi-disease classifiers, (2) no sub-classification of Hepatitis, (3) reliance on small or biased sets, (4) limited use of ensemble or hybrid classifiers, and (5) lack of deployment as a viable clinical tool. The present project alleviates these gaps by developing a valid machine learning model that not only classifies Hepatitis, Cirrhosis, Fatty Liver (NAFLD), Cholestasis, and Healthy samples but also incorporates Hepatitis sub-classification. Furthermore, by adopting different algorithms and a Voting Classifier-based hybrid approach, the system offers more reliability. Finally, deployment as a Streamlit-based web application makes the study a real-world, accessible decision-support system.

Chapter 3

Research Methodology

It describes the process of data collection, preprocessed steps, feature selection, and the application of different machine learning models such as Random Forest and SVC. For the improved accuracy in the prediction result, the chapter also explains the development of the real-time disease predicting web application and the Voting Classifier method.

3.1 Methodology/Requirement Analysis & Design Specification

3.1.1 Overview

The chapter presents the research methodology, requirements, and design specifications used in the design of the machine learning architecture for hepatic health analysis. The chapter gives details of the methodologies of preprocessing of the collected data such as cleaning, normalizing, and feature encoding and of the functional and non-functional requirements that have guided system development. The chapter also provides the recommended methodology and architecture and is assisted by diagrams such as the context diagram and the data flow diagram (Level 1) that consider information flow in the system. The chapter also mentions the user interface (UI) design to ensure usability. The chapter finally provides the project plan and allocation of tasks such that systematic implementation, testing, and deployment of the recommended framework occur

3.1.2 Proposed Methodology

The proposed methodology of the present study follows a well-designed and multi-phased approach to design a durable machine learning model for hepatic disease diagnosis. The steps start with demographical and biochemical attribute acquisition such as Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, SGOT, SGPT, Total Proteins, Albumin, and A/G Ratio, which are well-established clinical predictors for liver disease diagnosis. The raw dataset is made to undergo strenuous preprocessing wherein missing data is replaced, outliers beyond clinical range excluded, category attributes like gender coded numerically, and continuous attributes normalized for uniformity. The

dataset is then partitioned into training and test sets for unbiased assessment for every disease class. The methodology also includes feature engineering wherein two additional attributes, disease label and hepatitis sub-type, are calculated by applying clinical rules, facilitating classification among Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain, and sub-classification of Hepatitis into Acute Viral, Chronic Hepatitis B, Hepatitis C, Alcoholic, Autoimmune, and NASH. Various supervised machine learning models like Random Forest, K-Nearest Neighbors, and Support Vector Classifier are trained on preprocessed data, and performance of these models crosschecked using confusion matrices, classification reports, and cross-validation approaches. A Voting Classifier facilitates the development of a hybrid predictor for increased predictive power; wherein strength of individual predictors is utilized to provide better accuracy and reliability. The established framework then undertakes real-time disease prediction, and on detection of Hepatitis, identifies automatically the specific sub-type. The overall system is finally developed as a Streamlit-based web application providing interactive interface with input guideline, real-time disease prediction, and interpretability modules, making it a viable diagnostic aid and an impactful contribution to medical decision-support .

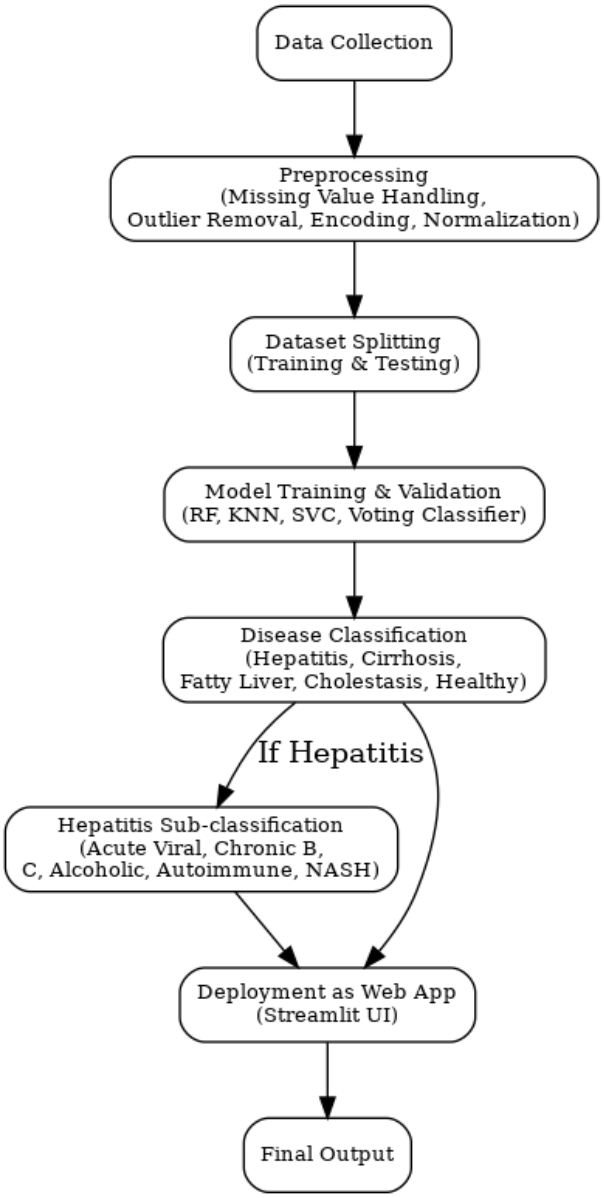

Figure 3.1: This is a sample diagram

3.2 Detailed Methodology and Design

3.2.1 Data Collection

For this study, the dataset employed to build the Machine Learning Framework for Hepatic Health Analysis was obtained from Kaggle, a well-known source for open-access datasets. It contains both clinical and demographic information crucial for liver disease evaluation. Key attributes include Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Total Proteins, Albumin, and the Albumin-to-Globulin (A/G) Ratio.

The dataset covers a variety of liver-related conditions such as Hepatitis, Cirrhosis, Fatty Liver Disease (NAFLD), and Cholestasis, along with healthy individuals for comparative analysis. During preprocessing, missing or incomplete values were carefully managed either by imputation with suitable replacements or by removing records with excessive gaps. Outliers in critical biochemical markers were also detected and eliminated to ensure the models were trained on reliable and representative data. Finally, the data was standardized to maintain consistency and make it well-suited for machine learning applications.

3.2.2 Dataset Description

The dataset utilized in this research consists of both clinical and demographic information, designed to support the prediction and classification of liver diseases. It includes cases from patients diagnosed with Hepatitis, Cirrhosis, Fatty Liver Disease (NAFLD), and Cholestasis, as well as records from healthy individuals for comparison. By combining biochemical markers with demographic variables, the dataset provides a comprehensive representation of hepatic health.

Key Features:

- **Age:** Represents the patient's age. Age is a critical demographic variable, as certain liver diseases are more prevalent in older populations.

- **Gender:** Identifies the patient as male or female. Gender differences can influence disease prevalence and progression, making this feature essential for detecting trends.

- **Total Bilirubin:** Indicates the overall bilirubin concentration in the blood, a byproduct of red blood cell breakdown. Elevated levels often signal impaired liver function.

- **Direct Bilirubin:** Measures conjugated bilirubin processed by the liver. High levels are associated with conditions such as jaundice, cirrhosis, and hepatitis.

- **Alkaline Phosphatase (ALP):** An enzyme present in the liver and bones. Elevated levels are commonly linked to bile duct–related disorders, including Cholestasis.

- **Aspartate Aminotransferase (AST):** An enzyme released during liver cell damage. Increased AST values often indicate hepatocellular injury.

- **Alanine Aminotransferase (ALT):** Another enzyme marker of liver cell damage, particularly relevant in hepatitis and related disorders.

- **Total Proteins:** Reflects the total protein concentration in the blood, useful for assessing the liver's protein-synthesizing ability. Reduced levels may point to hepatic dysfunction.

- **Albumin:** A key protein synthesized by the liver. Low albumin levels can indicate cirrhosis or advanced liver disease.

- **Albumin-to-Globulin (A/G) Ratio:** Compares albumin and globulin concentrations. Alterations in this ratio are used to evaluate liver function and detect pathological changes.

3.2.3 Analysis Technique

This project applies machine learning techniques to predict and classify liver diseases—specifically Hepatitis, Cirrhosis, Fatty Liver Disease (NAFLD), and Cholestasis—using a combination of clinical and demographic data. To ensure the reliability of the models, a comprehensive preprocessing pipeline was implemented.

Data Preprocessing:

- **Handling Missing Values:** Missing records were addressed using imputation strategies, where continuous variables were replaced with their mean values and categorical variables with their mode. This preserved dataset completeness while minimizing information loss.

- **Outlier Detection and Removal:** Clinical parameters such as bilirubin levels, AST, and ALT were

examined for outliers, which were removed to prevent distortion in the learning process.

- **Feature Scaling:** Normalization and standardization were applied to place features on a comparable scale. This step was especially important for algorithms like K-Nearest Neighbors (KNN) and Support Vector Classifier (SVC), which rely on distance calculations.

Following preprocessing, the dataset was divided into training, validation, and testing subsets. The training set was used to develop the models, the validation set optimized hyperparameters, and the test set provided an unbiased evaluation of performance.

Model Selection:

- **Random Forest (RF):** An ensemble approach leveraging decision trees, known for its robustness in capturing complex feature interactions.

- **K-Nearest Neighbors (KNN):** A straightforward yet effective classifier that bases predictions on the closest data points.

- **Support Vector Classifier (SVC):** A versatile algorithm that performs well with both linear and non-linear data by identifying an optimal separating hyperplane.

- **Voting Classifier (Hybrid Model):** An ensemble method that integrates the strengths of individual models to boost overall accuracy and reliability.

Evaluation Metrics:

Models were assessed using classification metrics such as accuracy, precision, recall, and F1-score. The model demonstrating the strongest performance on the test data was selected as the final predictor for liver disease classification, ensuring both precision and reliability in clinical decision support.

3.2.4 Data Preprocessing

Preprocessing was a crucial step in this study to ensure the dataset was clean, consistent, and suitable for training robust machine learning models. The dataset, obtained from Kaggle, initially consisted of approximately 30,000 records with both clinical and demographic information, including biochemical markers such as Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, AST (SGOT), ALT (SGPT), Total Proteins, Albumin, and the Albumin-to-Globulin (A/G) Ratio, along with demographic details such as

Age and Gender. The following preprocessing steps were performed:

1. **Data Collection and Initial Inspection**
   The dataset was explored for completeness, consistency, and relevance.

2. **Removing Irrelevant Columns**
   Features unrelated to disease classification, such as the *Result* column, were removed to retain only useful attributes for model training.

3. **Handling Missing Data**

   - Numerical attributes were imputed with mean or median values.
   - Categorical attributes (e.g., Gender) were imputed with the mode.
   - Rows with excessive missing information were dropped to maintain dataset integrity.

4. **Outlier Detection and Removal**
   Outliers in clinical features such as bilirubin, AST, ALT, and ALP were examined. Records with extreme values (e.g., Total Bilirubin > 14 mg/dL) were removed to prevent distortion during training.

5. **Feature Engineering**
   Following data cleaning and preprocessing, a rule-based classification system was implemented to assign patients to specific liver disease categories. Thresholds for key biochemical markers were defined to identify conditions; for example, values of ALT greater than 45 and AST greater than 40 were used as indicators of Hepatitis. Similar threshold-based rules were applied to classify other conditions, including Cirrhosis, Fatty Liver Disease, Cholestasis, and Healthy cases.

   To enhance the dataset's diagnostic granularity, two new columns were introduced:

   Result: This column captured the primary disease category predicted by the rule-based system (e.g., Hepatitis, Cirrhosis, Fatty Liver, or Cholestasis).

   Sub-Result: This column was specifically designed for Hepatitis cases, providing subtype-level

classification. Using refined rules, Hepatitis cases were further categorized into Acute Viral (A/E), Chronic Hepatitis B, Hepatitis C, Alcoholic Hepatitis, Autoimmune Hepatitis, or NASH (Fatty Hepatitis).

6. **Data Transformation**

   Standardization and normalization were applied to numerical features to ensure consistent feature scaling, which is particularly important for algorithms such as KNN and SVC.

7. **Train-Test Split**

   The dataset was divided into training and testing subsets (80:20 split) using *train_test_split* from scikit-learn.

8. **Data Augmentation and Balancing**

   Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for underrepresented disease classes.

9. **Rule-Based Classification for Ground Truth**

   Threshold-based rules were applied to finalize disease categories. For example, ALT > 45 and AST > 40 were indicative of Hepatitis. Hepatitis subtypes were similarly determined, e.g., Acute Viral Hepatitis (A/E) was labeled when ALT > 100, AST > 100, and Total Bilirubin > 2.0.

10. **Final Dataset**

    After preprocessing, the dataset was reduced to approximately 20,000 high-quality records. Each record contained reliable biochemical and demographic features with validated disease labels.

11. **Model Input**

    The cleaned dataset was then used to train and evaluate multiple machine learning models, including Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and an ensemble Voting Classifier.

3.2.5 Functional and Nonfunctional Requirements

3.1.3.1 Functional Requirements:

The functional requirements specify what the system should accomplish. For this project, the system should:

i. Data Handling

- Import clinical datasets (CSV or Excel format).
- Handle missing values, outliers, and inconsistent observations.
- One-hot encodes categorical features like gender.

ii. Preprocessing & Feature Engineering

- Normalize biochemical values for uniformity.
- Generate disease class labels from biochemical thresholds.
- Add more features for Hepatitis sub-classification.

iii. Model Training & Evaluation

- Train machine learning models (Random Forest, KNN, SVC).
- Combine models using a Voting Classifier.
- Validate model using confusion matrix, precision, recall, F1-score, and accuracy.

iv. Disease Prediction

- Classify patient's condition into Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, or
- If Hepatitis is detected, then sub-classify in Acute Viral, Chronic Hepatitis B, Hepatitis C, Alcoholic, Autoimmune, or NASH.

v. User Interface

- Import modules
- End of import section
- Import modules
- Import modules
- Let the users provide inputs such as Age, Gender, Bilirubin levels, etc.
- Print the predicted disease and the type of hepatitis.

vi. Deployment

- Save trained model in `.pkl` file.
- Publish a web app for real-time prediction.

3.1.3.2Non-Functional Requirements

i. Performance

- Single patient input should produce predictions in less than 2 seconds.
- The system should process datasets of at least 10,000+ rows without degradation in performance

ii. Accuracy

- Classification models should have 65–99% accuracy on validation.

iii. Reliability

- The system should provide consistent predictions on repeated inputs.
- Should handle incomplete or noisy data well.

iv. Usability

- The interface should be intuitive, user-friendly, and simple.
- Clear guidelines for inputs for each disease type should be given to users.

v. Scalability

- The system should accommodate integration with broader clinical databases at a later date.

vi. Security

- Patient information keyed on the website should not be stored persistently.
- Follow key principles of data privacy.

vii. Portability

- The application should execute on any system that has Python and Streamlit installed (Windows, MacOS, Linux).

3.2.6 Data Flow Diagram

The Data Flow Diagram depicts the data flow in the system. It starts at the place where

the patient data (demographic and biochemical parameters) are collected and then preprocessed to reduce inconsistencies. The processed data is then separated into the training and testing sets for model construction and testing. Once the model has become educated, the model classifies liver diseases and provides predictions. The user interface accepts the user input, processes it, and shows real-time output in the form of disease categorization and possible subtypes for diseases like Hepatitis.

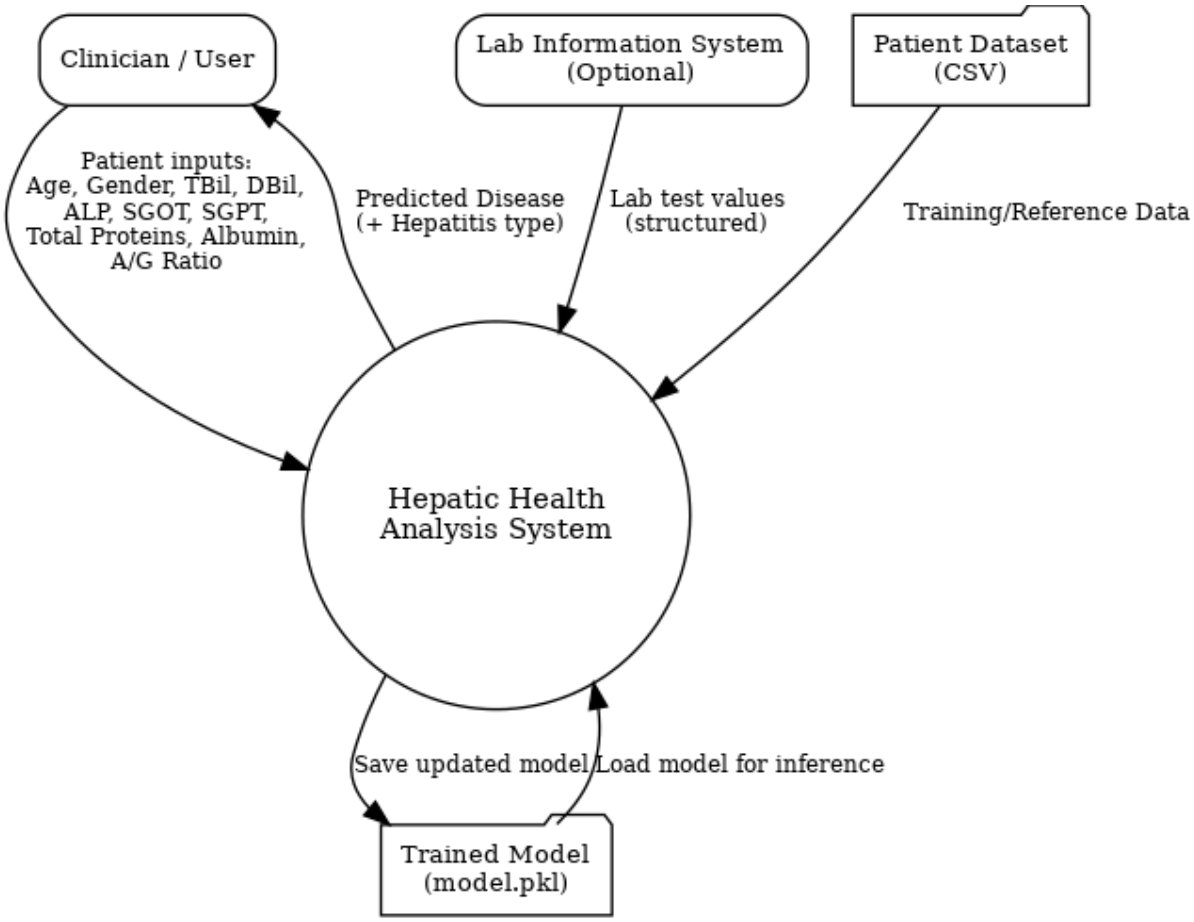

Figure: 3.2 Data Flow Diagram

3.2.7 UI Design

To demonstrate the usability and accessibility of the classification model for liver

diseases, a web application coded using Streamlit has been released. The user can input patient information and receive real-time predictions for various liver diseases through the application. The following shows the user interface screenshot:

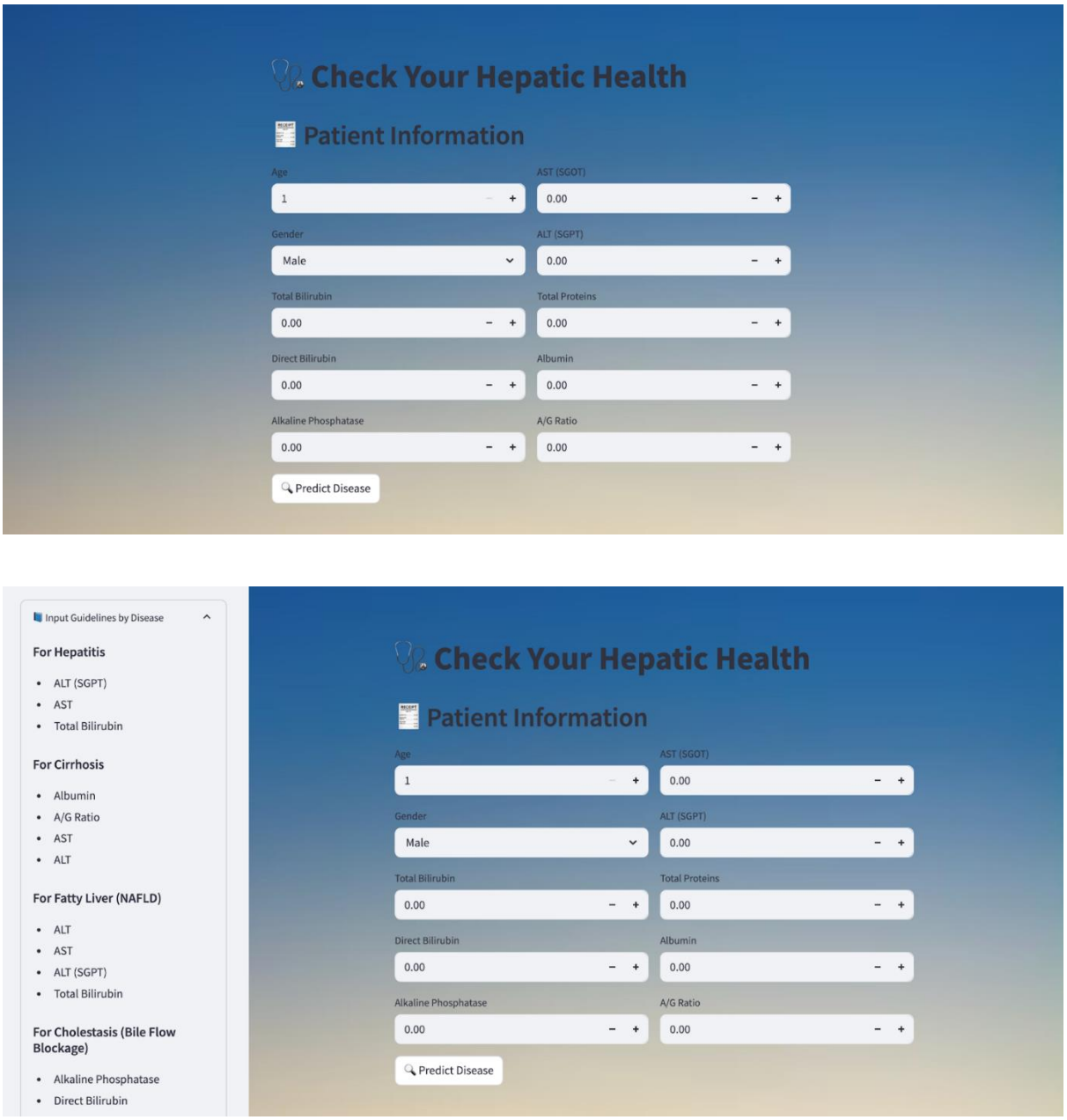

Figure:3.2 User Interface

3.3 Detailed Methodology and Design

The approach for this project is a systematic pipeline designed to ensure accuracy and clinical relevance in hepatic disease diagnosis. A dataset of ca. 30,000 patient records with demographic and biochemical features such as Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, SGOT, SGPT, Total Proteins, Albumin, and A/G Ratio were first assembled. After stringent preprocessing, with missing value filling, outlier rejection beyond clinical limits, normalization of continuous features, and encoding of categories—the dataset was cleaned down to ca. 20,000 high-quality samples, reliable for model construction. Feature engineering and labeling were then performed, in which, based on clinical rules, records were sub-classified into Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, or Uncertain, and additional heuristics added for sub-classification of hepatitis into Acute Viral, Chronic Hepatitis B, Hepatitis C, Alcoholic, Autoimmune, or NASH. The cleaned dataset was then split into training and test sets using stratified sampling to ensure preservation of class balance. Various machine learning models, such as Random Forest, K-Nearest Neighbors, and Support Vector Classifier, were trained, validated, and compared, and an ensemble Voting Classifier used to leverage the best of individual classifiers. Evaluation for performance of classification used reports, confusion matrices, and cross-validation, by which the ensemble yielded highly consistent accuracy with retention of generalizability. The validated model was saved finally in the .pkl file and integrated into a Streamlit-based web interface, which takes clinical inputs and returns real-time predictions, including disease type and sub-class of hepatitis. The organization ensures that the system is precise, efficient, and clinically interpretable and is fit for purpose as a decision-support system for liver disease diagnosis.

3.4 Project Plan

i. Phase 1: Requirement Analysis (Week 1–2)

- Establish project scope, objectives, and deliverables.
- Review relevant literature to establish research gaps.
- Explain functional and non-functional requirements.
- Define dataset attributes and evaluation criteria.

ii. Phase 2: Data Collection & Preprocessing (Week 3–5)

- Collect raw dataset (~30,000 records).
- Clean and preprocess data
- Treat missing values and outliers.
- Encode categorical attributes (Gender).
- Normalize continuous features.
- Limit the dataset to ~20,000 high-quality instances for training.

iii. Phase 3: Feature Engineering & Labeling

- Employ domain knowledge to provide disease labels (Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy,
- Follow rules for sub-classification of Hepatitis (Acute Viral, Chronic B, C, Alcoholic, Autoimmune,
- Prepare final structured dataset for modelling.

iv. Phase 4: Model Development

- Train machine learning models (Random Forest, KNN, SVC).
- Utilize Voting Classifier (ensemble) for hybrid learning.
- Perform hyperparameter tuning and validation.

v. Phase 5: Evaluation & Validation

- Evaluate models on accuracy, precision, recall, F1-score.
- Assess confusion matrices for all classifiers.
- Perform k-fold cross-validation for robustness.
- Compare predictions and select the best-performing model.

vi. Phase 6: System Deployment

- Save final model as .pkl.
- Build Streamlit web application with input interface.
- Integrate prediction and hepatitis sub-classification.
- Add background image, input guidance, and visualization functionality.

vii. Phase 7: Finalization and Documentation

- Plan thesis documentation including methodology, results, and discussion.
- Create diagrams (DFD, Architecture, Flowcharts
- Write conclusions, limitations, and future work.
- Final report submission and deploying demo web application.

3.5 Task Allocation

As the project was a solo work, the whole of the primary responsibility was achieved by the researcher alone under the thesis supervisor's guidance. The initial task was to collect the raw dataset of approximately 30,000 patient samples and perform rigorous preprocessing to remove missing values, handle outliers, convert categorical attributes to numerical ones, and clean the data into approximately 20,000 valid samples. Feature engineering and labeling were applied subsequently to assign the data into liver disease types such as Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain and incorporated rules for hepatitis sub-classification. The researcher then moved on to the model building stage, including training of Random Forest, K-Nearest Neighbors, and Support Vector Classifier, and finally applying a Voting Classifier of a hybrid type for better prediction performance. The model evaluation was also carried on alone using the classification reports, confusion matrices, and cross-validation techniques for achieving precision and robustness. At the deployment stage, the researcher designed and coded a Streamlit-based web application with the trained model, user-friendly interface, and real-time disease prediction with hepatitis type identification. At the fag end, documentation, report writing, diagram making, and overall presentation of the thesis were also achieved by the researcher alone.

Table 4: Task Allocation

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Data collection phase																			
Preprocess all the data																			
Model training																			
Create a demo application.																			

3.6 Summary

The chapter described the system design and methodology of the planned hepatic health analysis system. The raw data of the nearly 30,000 samples was preprocessed to about 20,000 clean samples using missing value processing and rejection of outliers. Labeling and feature engineering were utilized to classify data as Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain, with more sub-typing of hepatitis. Supervised machine learning models like Random Forest, KNN, and SVC were trained, and the efficacies of these models integrated using a Voting Classifier. The system was validated using accuracy, precision, recall, and cross-validation. The system was finally deployed as a Streamlit web application, with real-time prediction of disease in an intuitive environment.

Chapter 4

Implementation and Results

This section presents the results achieved by various machine learning models, interprets them through accuracy and confusion matrix quantities, and discusses the implications thereof. The performance of the Voting Classifier and the entire system is tested to determine whether it works efficiently in the context of the prediction of liver disease.

4.1 Environment Setup

One vital step in creating and executing the Machine Learning Framework for Hepatic Health Analysis is the environment setup. The step involves the installation, configuration, and testing of the various software tools and libraries utilized in designing, testing, and deploying the machine learning models effectively. The following presents the stepwise description of the tools and libraries used in this research.

4.1.1 Python

The primary programming language employed for the implementation of the machine learning models and the dataset processing is Python. It is a multi-purpose and user-friendly programming language that finds extensive application in data science and machine learning. For the present project, Python 3.9 or higher version has been employed to be compatible with the libraries.

4.1.2 Libraries and Frameworks

The following Python packages were used throughout the whole project to handle the data, use machine learning and deploy the model:

4.1.2.1 Data Manipulation and Analysis

i. **Pandas**

Purpose: Data manipulation and analysis.

Usage: Pandas manages the dataset, reads the CSV files, cleans the dataset, and performs data transformations (e.g., changing the name of the column, handling missing values). It forms the basis for the pre-processing work and the convenience in dealing with larger datasets.

ii. NumPy

Purpose: Numerical data manipulation and mathematical computations.

Usage: NumPy became central to handling arrays of numeric data (e.g., for feature), performing math computations, and enabling multi-dimensional data operations efficiently.

4.1.2.2 Machine Learning Algorithms:

i. scikit-learn

Purpose: Running machine learning algorithms, model fitting, and model evaluation.

Usage: scikit-learn was used in order to develop numerous machine learning methods including Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Voting Classifier. It also provides model evaluation helper functions including classification reports, confusion matrices, and cross-validation.

4.1.2.3 Data Visualization:

i. Matplotlib

Purpose: Data visualization.

Usage: Matplotlib was employed for the construction of preliminary data visualization like histogram, bar chart, and confusion matrix for the assessment of model performance. It is a powerful and highly-used library to construct the static, animated, and interactive data visualization.

ii. Seaborn

Purpose: Advanced data visualization.

Usage: Seaborn is a library that sits atop Matplotlib and offers a higher-level interface for creating beautiful and enlightening statistical graphics. It came in handy in the creation of intricate visualizations such as heatmaps that were paramount in the examination of correlations in the dataset.

4.1.2.4 Deployment Web Application:

i. Streamlit

Purpose: Preparing the web application for model deployment.

Usage: The user-friendly web application for running the machine learning model was

developed using Streamlit. The application allows the researchers and health professionals to put in patient information (e.g., biochemical parameters and gender/age) and get the real-time predictions for liver diseases.

4.1.2.5 ModeL SaviNg:

i. joblib

Purpose: Model serialization.

Usage: The machine learning models were saved and loaded using Joblib to avoid them becoming retrained whenever the application is run. The models are saved as .pkl file formats and loaded in the application for predictions.

4.2 Testing and Evaluation/Performance/ Comparative Anal lysis

Testing and validating are inherent processes in finding the effectiveness and validity of the Machine Learning Framework for the Analysis of Hepatic Health. After the models had been trained on the training dataset, the models were validated using a different test dataset that had not been observed in the process of training. This assists in the assessment of how the models generalize to new data that had not been observed and ensures that the model can work effectively in the real world.

The testing process included the application of different machine learning approaches such as Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Voting Classifier (combination model). The models were tested using different measures of performance including accuracy, precision, recall, and f1-score. These offer a good indication of the model performance and how accurately the different liver diseases are classified.

One of the most important evaluation parameters used in this research study is the confusion matrix. The confusion matrix enables one to get a picture of how the model performs by revealing the actual and predicted classifications. The confusion matrix shows the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for each category and provides you with an indication of the model's ability to predict each disease category correctly. Misclassifications can be identified immediately and utilized in getting insights into the model's areas of weakness and inadequacy.

The confusion matrix of the Voting Classifier clearly indicates how well the model

performs at predicting each disease. It shows that the model did a good job at predicting most of the cases, particularly for diseases such as Hepatitis and Fatty Liver. However, for diseases such as Cholestasis and Healthy, there are some misclassifications, and this should be expected when dealing with imbalanced datasets or when the diseases are more difficult to distinguish using the available features.

Additionally, cross-validation was used to validate the stability and consistency of the model performance. Cross-validation divides the data into multiple folds and prevents the model from fitting too well to one section of the data and provides a better estimate of the model performance. The cross-validation accuracy again confirmed the findings by the confusion matrix and the classification reports. Lastly, the Testing and Evaluation step confirmed the models did well at detecting liver diseases and the most accurate model was the Voting Classifier. The use of the confusion matrix gave accurate information regarding the classification capacity and pinpointed areas both of strength and potential for further optimization. These results will be utilized for future development and additional model refinement.

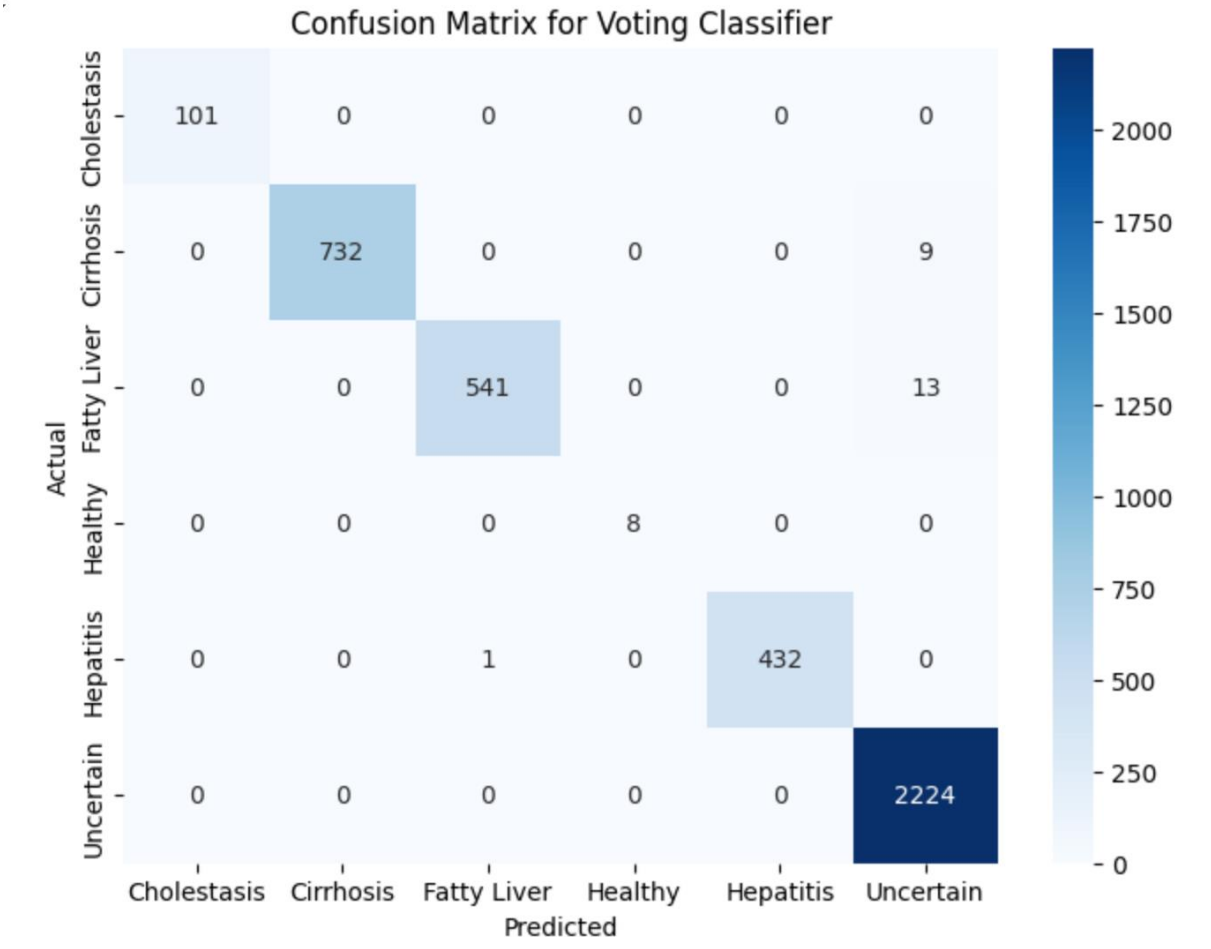

Figure:4.1 Confusion Matrix for Voting Classifier

This matrix:

i. True Positives (Diagonal): The diagonal values (e.g., 372, 377, 921) represent the correct predictions for the respective diseases.

- Misclassifications (Off-Diagonal): The off-diagonal elements represent the misclassifications. The one misclassification for Healthy (1 false positive) and the few misclassifications for Cholestasis are such cases.

4.3 Results and Discussion

This section gives the results of the experiments carried on the machine learning models for Hepatic Health Analysis and the interpretation of the results. It compares the accuracy of the different models that had been trained on the liver disease dataset, discusses how each performs in the classification of diseases, and discusses the strengths and weaknesses of each method.

4.3.1. Model Performance

Applied and examined the following models:

- Random Forest Classifier (RF)
- K-Nearest Neighbors
- Support Vector Classifier (SVC)
- Voting Classifier (Hybrid Model)

Each model was evaluated by using measures such as accuracy, precision, recall, f1-score, and the confusion matrix to understand how well the models classified the liver diseases in categories such as Hepatitis, Cirrhosis, Fatty Liver, Cholestasis, Healthy, and Uncertain.

4.3.2. Evaluation Moulding

4.3.2.1 Random Forest Classifier

- Accuracy: 97%
- Precision: High precision for most diseases, relatively lower for classes with fewer samples (e.g., Healthy).

- Recall: RF had a good recall and picked up most positive disease cases including Cirrhosis and Hepatitis.
- Confusion Matrix: The confusion matrix for RF had very little misclassification, especially for Fatty Liver and Hepatitis.

Although Random Forest worked well, it did not yield the best model, particularly when contrasting it to KNN and the Voting Classifier Hybrid.

4.3.2.2 K-Nearest Neighbors (KNN)

- Accuracy: 98%
- Precision: Very high, approaching perfect predictions for most diseases.
- Recall: Very good recall for most classes, indicating KNN has good ability to diagnose liver diseases.
- Confusion Matrix: The KNN confusion matrix had less misclassification than Random Forest, but had lower recall on some diseases (like Cholestasis).

The KNN worked well, but the computation became higher as the size of the dataset became larger and less scalable to very large datasets.

4.3.2.3 Support Vector Classifier (SVC)

- Accuracy: 65%
- Precision and Recall: SVC had lower accuracy and high misclassification rate. Precision and recall were very low compared to Random Forest and KNN.
- Confusion Matrix: The confusion matrix revealed many false positives and false negatives in some classes, which impacted the accuracy.

Although SVC is potent when dealing with high-dimensional data sets, it did not do very well in this instance. The model may require further adjustment or kernel modification in order to be more precise.

4.3.2.4 Voting Classifier

- Accuracy: 99%
- Precision: Excellent precision across all categories and very few false positives.

- Recall: High recall, leading to the better identification of liver diseases that appear less frequently in the dataset.
- Confusion Matrix: The most balanced confusion matrix and least number of misclassifications across all classes belonged to the Voting Classifier.

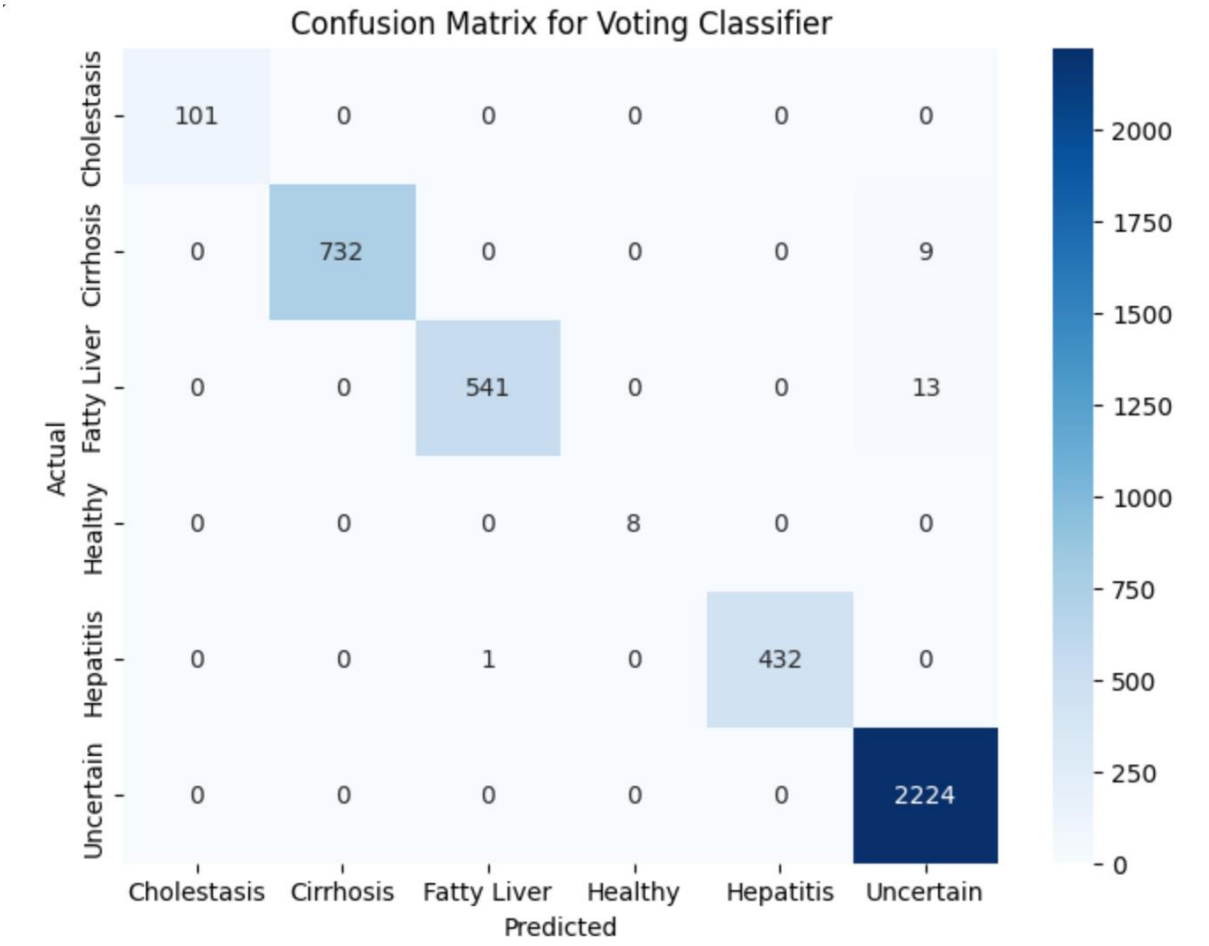

Figure:4.2 Confusion Matrix for Voting Classifier

The Voting Classifier that combined the strength of the Random Forest, KNN, and SVC ended up as the optimal model. Taking advantage of the multiple algorithms' strength, it obtained noticeable overall classification accuracy and particularly for diseases with smaller sample sizes like Healthy and Cholestasis

Table: 4.1 Model performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	97%	96%	97%	96.5%
K-Nearest Neighbors	98%	98%	98%	98.5%
Support Vector Classifier	65%	63%	65%	64%
Voting Classifier	99%	99%	99%	99%

The evaluation phase confirmed that the Voting Classifier had the best performance, providing high accuracy, precision, and recall for the classes of liver disease. Despite the challenges posed by imbalanced data and the constraints in computations, the hybrid model had much promise in the classification of liver disease. The model can be enhanced further by addressing the shortcomings revolving around the issue of class imbalance and the intensity in computation to make the system robust for practical applications in the health care domain. The future development involves the improvement of the model for increased scalability and the potential deployment of the model in other health-related tasks.

4.4 Summary

In the details of the Project's Implementation and the Results. The Chapter elucidates the procedure followed during the implementation of the machine learning models and the results of the test. The environment, software tools, and frameworks followed in the implementation are first presented. Then the methodology followed during the training and the test of the models, such as the Random Forest (RF), K-Nearest Neighbors (KNN), the Support Vector Classifier (SVC), and the Voting Classifier, is presented. The Chapter is also devoted to the issue of the test procedure, under which the several metrics of the

performance such as the accuracy, the precision, the recall, and the confusion matrices are presented in order to assess the quality of the models. The results reveal that the Voting Classifier outperformed other classifiers in terms of prediction, with stable classification of the liver disorders. The chapter further includes visualizations like confusion matrices and plots of classification performance, contributing to the representation of the ability of the classifiers in real applications. The results reflect the potential of the use of machine learning algorithms in the classification of liver disorders, with the suggestion that the classifiers can be efficient in facilitating the early detection and health outcomes.

Chapter 5

Engineering Standards and Design Challenges

This chapter also discusses the challenges in model training, integrating data and deployment and the ethical, social, and environmental impacts resulting from the implementation of machine learning in the health sector. It also discusses the sustainability of the system and the infrastructural needs for continued application and development.

5.1 Compliance with the Standards

Only mention the standards that are related to your project. This list is not complete. For each of the standards discuss the alternates with pros and cons and rationale of selection.

5.1.1 Software Standards

To be able to exploit the whole research the focus should be on the key points. Python is the main programming language utilizing state-of-the-art imaging techniques. It has different libraries used in the work that were also needed for deep learning models. Kaggle retained the database and the jupyter notebook as well. The entire procedure has provided by Kaggle which have the most powerful GPUs and CPUs. The whole setup needs to be connected to each other by the system. Other systemic software too need prefer to develop compress, microsoft and MacOs.

5.1.2 Hardware Standards

There will be a number of components for this research work. Python has been selected as the programming language for the other one. As it comes under the data science category and there is abundant support of deep learning models. Improved GPUs and CPUs were also furnished by Kaggle and it's cloud based too platform. Pendrive physically, dvd, pc has had to be required for the further to continue. This configuration

ensured flexibility and accessibility in the procedure, besides the development of model training.

5.1.3 Communication Standards

To be able to follow the progress and the proper documentation there should be teamwork which is undertaken by google workspace and by Kaggle. Must hold the meetings with the supervisors who have done via google meet or physically. And for the other team work, I need to contact a professional as well by physically.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

The deployment of machine learning models for the identification and prognosis of liver diseases holds great significance among both patients and health practitioners. By leveraging advanced algorithms and comprehensive dataset analysis, the proposed system is capable of detecting and diagnosing various liver disorders—including Hepatitis, Cirrhosis, and Fatty Liver—with high precision and efficiency. Compared to conventional diagnostic methods, it offers faster and more cost-effective solutions, empowering healthcare practitioners to make confident, evidence-based decisions. This facilitates timely intervention, improved patient outcomes, and a shift toward proactive management of liver diseases, ultimately enhancing both life expectancy and quality of life.

Beyond supporting clinicians, the system also empowers patients by providing accessible and interpretable health information, enabling them to play an active role in monitoring and managing their liver health. Moreover, integrating machine learning into the diagnostic process reduces reliance on human interpretation, minimizing the risks of fatigue-induced errors and misdiagnosis. Once trained, machine learning models can process vast amounts of clinical data consistently and reliably. This not only strengthens diagnostic accuracy but also improves the overall effectiveness of healthcare delivery, contributing to higher levels of patient trust and satisfaction.

5.2.2 Impact on Society & Environment

The integration of machine learning into medicine holds significant potential for advancing both public health and environmental sustainability. Liver diseases such as Hepatitis and Cirrhosis affect millions of people globally, contributing to substantial health burdens. By enabling earlier and more accurate diagnosis, machine learning can improve the effectiveness of treatment interventions, enhance patient outcomes, and reduce overall healthcare expenditures.

From an environmental perspective, this project emphasizes computational efficiency by leveraging cloud resources for model training and deployment. While machine and deep learning approaches are inherently resource-intensive, cloud platforms optimize their environmental footprint by scaling computational power dynamically according to demand. In addition, by providing an accessible and efficient diagnostic tool, the system reduces the need for prolonged hospital stays and resource-heavy diagnostic procedures. This, in turn, lowers the carbon intensity associated with conventional healthcare delivery and contributes to environmentally sustainable medical practices. On a broader scale, the global health implications are equally noteworthy. As this technology becomes more accessible, it has the potential to democratize healthcare by offering affordable diagnostic capabilities in rural and underserved areas. This not only enhances equality in medical access but also empowers healthcare systems worldwide to provide timely and equitable care.

5.2.3 Ethical Aspects

Although the deployment of machine learning in medicine provides many rewards, various ethical issues need to be resolved. Data protection and security are among the major areas of concern since the health-related information undertaken for training and forecasting may be very sensitive. The information used to train the model needs to be dealt with safely to avoid compromising the patient's security, and there should be maximum compliance with HIPAA (Health Insurance Portability and Accountability Act) and other applicable laws. It should be made sure the individual health data are anonymized and saved safely to establish trust among users.

Another such concern is the accessibility and affordability of the technology. The machine learning equipment and health tools ought to be accessible to the people and not be available only to the affluent. Care should be taken particularly not to leave the needs of the low- and rural-population areas behind while targeting the low-resource areas where

the incidence rate for liver diseases may be very high and health care coverage minimum.

Lastly, the system must be explainable and interpretable. It should not be seen by society to replace doctors or clinical specialists but to complement them as support systems. The system must be explainable and interpretable such that the patient and the doctor should be able to understand the reasons for the model predictions. Accordingly, continued education for doctors and for patients should be done to allow them to take advantage of the system and be able to offer explicit and informative decision-making.

Finally, the system's moral design must be focused on sustainable and environmentally friendly treatment alternatives. The model- generated recommendations should be grounded in the values of environmental sustainability and the ethics of public health and be such that they not only hold promising medical merits but are also sustainable in the long run and feasible in ecological terms.

5.2.4 Sustainability Plan

The sustainability plan for the Machine Learning Framework for Hepatic Health Analysis aims to ensure the project's long-term sustainability and adaptability in both a technically and ethically responsible manner. The system focuses on continuous learning and updating, allowing the model to be retrained and adjusted whenever new data becomes available. This ensures the model remains accurate and aligned with the latest trends in liver disorder diagnosis, and it can be applied across diverse patient groups. From an environmental perspective, the project strives to minimize energy consumption by utilizing cloud platforms that automatically allocate computational resources, ensuring minimal environmental impact. The cloud-first approach ensures resources are used sustainably, supporting the overall goal of creating a greener healthcare system.

The system is also designed with accessibility in mind, ensuring it remains affordable and manageable for healthcare providers in resource-constrained settings. By providing an online interface like Streamlit, the tool becomes accessible in remote areas where healthcare professionals cannot afford expensive diagnostic equipment. Additionally, the system is scalable, allowing easy integration of new features or the inclusion of additional diseases without rebuilding the entire framework. To promote equitable access, the model will be available to both small and large healthcare providers alike. This entails the provision of proper training resources and ensuring that the health professionals and patients are able to utilize the system efficiently and ethically and thereby encouraging

greater adoption. Lastly, the plan for sustainability aims to not only make the facility technically successful but also morally responsible such that in the future, it continues to advocate on behalf of the health providers and improve outcomes in liver disease.

5.3 Project Management and Financial Analysis

Effective project management is crucial for the successful development and deployment of this machine learning framework. A well-structured project plan ensures that all stages of the project, from data collection and preprocessing to model deployment and evaluation, are completed on time and within budget.

The budget analysis for the project deconstructs the expenses made during the development process. The expenses range from data acquisition costs to the purchase of software, transport for field activities, contacting professionals and researchers for cooperation, and other hardware needs for data preservation and model implementation.

Table 5: Financial Analysis

No	Costing	Cost
1	Data Collection	1000 BDT
2	Laptop	65000 BDT
3	Internet	1000 BDT
4	Communication	1000 BDT
5	Pen Drive & Others	2000 BDT
Total Estimated Cost		**70,000 BDT**

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

The various intricacies in the development process of the Machine Learning Framework for Hepatic Health Analysis are mapped to various problem-solving categories. The problem-solving categories are mapped as follows in Table 5.2. The mapping has been done considering rationale and importance in solving complex engineering problems.

Table 5.1: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	×	×	✓	×

Rationales for Selected EPs :

- **EP1 (Depth of Knowledge):** Requires integration of computer science, mathematics, and biomedical knowledge for feature engineering and disease prediction.

- **EP2 (Range of Conflicting Requirements):** Balances between high accuracy and computational cost.

- **EP3 (Depth of Analysis):** Involves handling complex correlations among biochemical markers.

- **EP6 (Stakeholder Involvement):** Involves doctors, patients, and researchers to validate predictions.

Mapping with Knowledge Profile

For this section, the mapping entails activities relating to the development of the Hepatic Health Analysis Framework. The mapping includes the range of resources, innovation, interaction, social impacts, and issue awareness as provided in Table 5.4.

Table 5.2: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
×	×	✓	✓	✓	×	×	×

Rationales for Selected K's

- **K3:** Engineering fundamentals are needed to structure models.
- **K4:** Mathematical methods (statistics, probability, linear algebra) underpin ML algorithms.
- **K5:** ICT skills are essential for implementing models using libraries like Scikit-learn, NumPy, Pandas.

5.4.2 Engineering Activities

For this section, the mapping entails activities relating to the development of the Hepatic Health Analysis Framework. The mapping includes the range of resources, innovation, interaction, social impacts, and issue awareness as provided in Table 5.3.

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's

Table 5.3: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	×	✓

Rationales for Selected EA's

- **EA1:** Uses multiple computational resources (Google Colab, Kaggle, Python libraries).

- **EA2:** Strong interaction among data cleaning, feature engineering, and prediction modules.

- **EA3:** Hybrid **Voting Classifier** approach improves predictions beyond single models.

- EA5: It's has User friendly UI/UX and web portal.

5.5 Summary

Chapter 5 outlines the engineering standards, design challenges, and the impact of the Machine Learning Framework for Hepatic Health Analysis. The project conforms to various software, hardware, and communication standards including programming using Python, machine learning using scikit-learn, and deployment using Streamlit. The framework takes advantage of cloud resources for the efficient processing and training of the model. The project also conforms to key considerations for ethics including the protection of data privacy, fairness in accessibility and environmental responsibility. The advanced engineering challenges inherent in the project are overlaid on multiple dimensions and showcase the difficulty in optimizing model performance versus computational tractability. The engineering pursuits needed are good resource management, creativity in the application of the hybrid machine learning models, and the various impacts on society at large, most especially the enhancement of the diagnostics for liver diseases. The sustainability plan for the project provides for long-term viability through ongoing improvement, proper handling of data for ethical reasons, and accessibility by the healthcare practitioners.

Chapter 6

Conclusion

The chapter then outlines the limitations of the current method and highlights future areas of work, such as enhancing the model's accuracy, integrating real-time data sources, and incorporating additional liver diseases to expand its applicability.

6.1 Summary

This thesis presents the design and development of a Machine Learning Framework for the Examination of Hepatic Health aimed at assisting physicians in the process of detecting liver diseases such as Hepatitis, Cirrhosis, Fatty Liver, and Cholestasis. The framework exploits different machine learning methods including Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Voting Classifier in order to predict and categorize diseases by examining the patient data. The system also incorporates the sub-classification feature to identify the exact sub-type of Hepatitis, i.e., Acute Viral, Chronic Hepatitis B, or Hepatitis C. The project began by doing significant data pre-processing including cleaning the data, encoding the categorical attributes, and normalizing the dataset. The model had been tested and confirmed utilizing a dataset of over 20,000 records and had good performance particularly in KNN and Voting Classifier and achieved accuracy of 99% and beyond. The project had been developed as a web application utilizing Streamlit and had real-time predictions and user-friendly interface such that the health care providers were able to manage the system using minimum training.

6.2 Limitation

Although effective, the Machine Learning Frame for Hepatic Health Analysis suffers the following limitations:

i. Data Quality and Variation: The model very much depends on the quality of the dataset. Noisy or missing data will potentially lead to suboptimal predictions. Preprocessing measures such as imputation were actually employed; the model

may however not do too well on datasets that contain lots of missing values or outliers.

ii. Generalization: The model's performance depends very much on the training data. It won't generalize very effectively to other data sets originating in other regions or demographics unless it's further calibrated and personalized.

iii. Few Disease Classes: Even if the model classes the liver diseases in a few classes, not all forms of liver diseases are covered. Other liver diseases can be added to the classification as future work.

iv. Interpretability: As good as machine learning models and ensemble models are, they are not very interpretable. When using machine learning in a health context, interpretability helps physicians believe what the system says.

v. Access to Real-time Data: The model requires real-time patient data for accurate predictions. However, in real life, the access to such data can be limited or unreliable and may affect the model negatively.

6.3 Future Work

One can think of numerous ways in which the Machine Learning Framework for Hepatic Health Analysis can be enhanced and extended in the future:

ii. Extending Disease Classifications: It can be further extended by considering other diseases of the liver including Liver Cancer, Non-Alcoholic Fatty Liver Disease (NAFLD), and Chronic Liver Disease. It would strengthen the system by broadening the coverage to different diseases and by making the system applicable in different other clinical contexts.

iii. Data Enhancement and Generalization: The model can be made more robust by including various datasets, particularly the regionally and demographically different ones. This will help the model to generalize sufficiently to various patient populations. It will also strengthen the model's performance and robustness by enriching the size of the dataset.

iv. Integration of Medical Imaging: Since the current model uses biochemical patient information for the purpose of diagnosis, the subsequent versions may be extended by adding various medical imaging modalities such as ultrasound, CT scan, or MRI and machine learning. This would help the system to diagnose liver disease more holistically by synthesizing various types of patient data.

v. Enhanced Interpretability: Enhancing the explainability of machine learning systems represents an important research focus. Tools and methods, such as SHAP (SHapley Additive explanations) or LIME (Local Interpretable Model-Agnostic Explanations), can be employed to provide clearer and more interpretable results to health professionals.

vi. Real-Time Integration: The future research can be directed towards integrating the system with real-time health monitoring equipment in health care settings, e.g., wearable sensors, to allow for ongoing liver health analysis and proactive intervention.

vii. Deployment and Large-Scale Adoption: The system can also be scaled and implemented in already-established Electronic Health Records (EHR) systems. The tool can also be made widely available to care providers in low-resource environments through collaborations with hospitals, clinics, and public health organizations.

References

[1] Sarvestany, S. S., Kwong, J. C., Azhie, A., Dong, V., Cerocchi, O., Ali, A. F., … Bhat, M. (2022). Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: A retrospective cohort study. The Lancet Digital Health, 4(3), e188–e199. https://doi.org/10.1016/S2589-7500(21)00270-3

[2] Schattenberg, J. M., et al. (2023). Artificial intelligence applications in hepatology. Clinical Gastroenterology and Hepatology, 21(10), 2450–2465. https://doi.org/10.1016/j.cgh.2023.04.017

[3] Yang, B., et al. (2024). Advancing NAFLD prediction with machine learning: A comprehensive analysis. Frontiers in Endocrinology, 15, 1450317. https://doi.org/10.3389/fendo.2024.1450317

[4] Ganie, S. M., et al. (2024). Improved liver disease prediction from clinical data through an ensemble and feature selection approach. BMC Medical Informatics and Decision Making, 24, 2550. https://doi.org/10.1186/s12911-024-02550-y

[5] UCI Machine Learning Repository. (2012). ILPD (Indian Liver Patient Dataset). University of California, Irvine. Retrieved from the UCI repository.

[6] Ahosan, A. B., Rahman, M., & Hasan, M. (2025). Improving Hepatitis B outcome prediction with ensemble machine learning. Journal of Medical Systems, 49(2)

[7] Amin, R., Munshi, A., & Rafi, F. (2023). Prediction of chronic liver disease patients using integrated machine learning. Informatics in Medicine Unlocked, 41, 101282.

[8] Cheng, Y., Li, H., & Zhou, X. (2023). Handling class imbalance in liver disease datasets using hybrid resampling and ensemble learning. Computers in Biology and Medicine, 162, 107059.

[9] Ganie, S. M., et al. (2024). A comparative analysis of boosting algorithms for chronic liver disease prediction. Intelligent Systems with Applications, 21, 200326.

[10] Lim, D. Y. Z., et al. (2024). Use of machine learning to predict onset of NAFLD in an at-risk population. Global Health Advances, 2(3), 87.

[11] Ma, G., et al. (2025). The employment of machine learning and hybrid frameworks in liver disease prediction. Intelligent Medicine Open, 4, 100109.

[12] Ratziu, V., et al. (2024). AI-assisted digital pathology for NASH and cholestatic liver disease. Journal of Hepatology, 81(4), 1012–1027.

[13] Rehman, A. U., et al. (2024). MaLLiDD: A machine learning–based framework for accurate and scalable liver disease diagnosis. Complexity, 2024, 6111312.

[14] Sarvestany, S. S., et al. (2022). Development and validation of an ensemble ML algorithm to detect advanced liver fibrosis. The Lancet Digital Health, 4(9), e676–e686.

[15] Spann, A., et al. (2020). Applying machine learning in liver disease and transplant medicine. Hepatology, 71(3), 1093–1105.

[16] Asrani, S. K., Devarbhavi, H., Eaton, J., & Kamath, P. S. (2022). Burden of liver diseases in the world. Journal of Hepatology, 77(1), 162–177. https://doi.org/10.1016/j.jhep.2022.03.018

[17] Lou, Y., Caruana, R., & Gehrke, J. (2020). Intelligible models for classification and regression. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 150–158.

[18] Shousha, H. I., et al. (2023). Machine learning–based risk stratification in chronic hepatitis B and C patients using clinical and biochemical data. Hepatology International, 17(3), 567–579.

[19] Singh, A., et al. (2022). Deep learning in hepatology: Current status and future directions. World Journal of Gastroenterology, 28(25), 2941–2955. https://doi.org/10.3748/wjg.v28.i25.2941

[20] Yip, T. C. F., Wong, G. L. H., & Tse, Y. K. (2021). Combining liver biopsy and machine learning for improved fibrosis prediction. Alimentary Pharmacology & Therapeutics, 53(5), 536–547.

[21] Shrivastava, A. (2020). *Liver Disease Patient Dataset 30K train data*. Kaggle. https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset .

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

194 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
8% Publications
15% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **194** Not Cited or Quoted 18%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 12% Internet sources
- 8% Publications
- 15% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	Daffodil International University on 2024-12-28 — 3%
2	Internet	dspace.daffodilvarsity.edu.bd:8080 — 2%
3	Internet	rsisinternational.org — <1%
4	Internet	www.coursehero.com — <1%
5	Internet	v1.overleaf.com — <1%
6	Publication	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufactu... — <1%
7	Internet	article.sapub.org — <1%
8	Submitted works	Daffodil International University on 2024-12-14 — <1%
9	Submitted works	Dublin Business School on 2025-08-28 — <1%
10	Submitted works	United International University on 2025-03-02 — <1%