

Explainable Fusion Model for Non-Alcoholic Fatty Liver Disease Risk Prediction

by

Ayush Hassan Raiyan (223014015)

Jarin Tasnim Eara (223014136)

Safaina Khan Oishi (223014144)

Eshrat Kamal Nova (223014197)

*Capstone project report (CSE 4098C) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Dr Muhammad Golam Kibria, PhD



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF LIBERAL ARTS BANGLADESH

SPRING 2026

© *Ayush Hassan Raiyan, Jarin Tasnim Eara, Safaina Khan Oishi and Eshrat Kamal Nova*
All rights reserved

DECLARATION

Project Title Explainable Fusion Model for Non-Alcoholic Fatty Liver
Disease Risk Prediction
Authors *Ayush Hassan Raiyan, Jarin Tasnim Eara, Safaina Khan Oishi*
and *Eshrat Kamal Nova*
Student IDs 223014015, 223014136, 223014144 and 223014197
Supervisor Dr Muhammad Golam Kibria, PhD

We declare that this capstone project report entitled *Explainable Fusion Model for Non-Alcoholic Fatty Liver Disease Risk Prediction* is the result of our own work except as cited in the references. The capstone project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Ayush Hassan Raiyan
223014015

Department of Computer Science and
Engineering
University of Liberal Arts Bangladesh

Jarin Tasnim Eara
223014136

Department of Computer Science and
Engineering
University of Liberal Arts Bangladesh

Safaina Khan Oishi
223014144

Department of Computer Science and
Engineering
University of Liberal Arts Bangladesh

Eshrat Kamal Nova
223014197

Department of Computer Science and
Engineering
University of Liberal Arts Bangladesh

Date: May 3, 2026



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the capstone project report entitled **Explainable Fusion Model for Non-Alcoholic Fatty Liver Disease Risk Prediction**, submitted by **Ayush Hassan Raiyan** (Student ID: 223014015), **Jarin Tasnim Eara** (Student ID: 223014136), **Safaina Khan Oishi** (Student ID: 223014144) and **Eshrat Kamal Nova** (Student ID: 223014197) are undergraduate students of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Dr Muhammad Golam Kibria, PhD
Professor & Head

Suravi Akhter
Lecturer

Jannatul Ferdous Ruma
Senior Lecturer

Place: Dhaka

Date: May 3, 2026

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, *Dr Muhammad Golam Kibria, PhD*, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Ayush Hassan Raiyan, Jarin Tasnim Eara, Safaina Khan Oishi and Eshrat Kamal Nova
University of Liberal Arts Bangladesh

Date: May 3, 2026

Dedicated to
my loving Mother

– *Ayush Hassan Raiyan*

To my beloved parents
the source of my inspiration

– *Jarin Tasnim Eara*

Dedicated to my
teammates
*for the laughter and constant
encouragement that carried
us through*

– *Safaina Khan Oishi*

To my cherished parents
*for their unwavering belief in
me*

– *Eshrat Kamal Nova*

Contents

1	Introduction	1
1.1	Problem statement	2
1.2	Aims, objectives and Motivation	3
1.2.1	Aims	3
1.2.2	Objective	3
1.2.3	Motivation	4
1.3	Project Specification	4
1.3.1	Input Features	4
1.3.2	Prediction Targets	5
1.3.3	Performance Goals	5
1.4	Expected Outcomes	5
1.4.1	Model Training Outcomes	5
1.4.2	Explainability Outcomes	5
1.4.3	System Outcomes	5
2	Literature Review	6
2.1	NAFLD: Epidemiology, Risk Factors, and Diagnostic Challenges . .	6
2.2	Interpretability in NAFLD AI and ML Research	7
2.2.1	Background and Importance	7
2.2.2	Key Interpretability Approaches in Reviewed Studies	7
2.2.3	Trends and Limitations	8
2.3	Machine Learning in NAFLD Prediction	8
2.3.1	Overview of ML Applications	8
2.3.2	Commonly Used Algorithms	9
2.3.3	Performance Metrics in Literature	9
2.3.4	Challenges and Limitations	10

2.3.5	Identified Gaps	10
2.4	Handling Missing Data using Matrix Decomposition Techniques . .	11
2.5	Dataset Summary in Classical NAFLD Studies	12
2.6	Machine Learning–Based NAFLD Prediction Literature Analysis . .	14
2.7	Matrix Decomposition-Based Literature Analysis	28
2.8	Dataset Analysis Table	30
3	Proposed System for Non-Alcoholic Fatty Liver Disease Risk Prediction	31
3.1	System Overview	31
3.2	Methodology	32
3.3	Requirement Analysis	34
3.3.1	Functional Requirements	34
3.3.2	Software and Platform Requirements	34
3.3.3	Hardware and Runtime Configuration	35
4	Design and Development	36
4.1	Dataset	36
4.1.1	Collected Clinical Dataset	36
4.1.2	Public NAFLD Datasets	37
4.2	Dataset Preprocessing and Feature Alignment	39
4.3	Matrix Decomposition for Missing Value Completion	39
4.4	Class Imbalance Handling and Dataset Splitting	40
4.5	Model Training and Ensemble Learning	40
4.6	Explainability Using SHAP	40
4.7	Web Application Overview	41
5	Result Analysis	42
5.1	Model Performance Evaluation	42
5.2	SHAP Explainability	50
6	Web Application	55
6.1	Web Application Implementation	55
6.2	Model Deployment	56
6.3	User Interface and Functionality	56
6.4	Website Demonstration	57

7	Conclusions	66
7.1	Social, Legal, Ethical, and Environmental Issues	66
7.1.1	Social Issues	66
7.1.2	Legal Issues	66
7.1.3	Ethical Issues	66
7.1.4	Environmental Issues	67
7.2	Brief Summary	67
7.3	Group Members Contribution	67
7.4	Future works	68
	References	69

List of Figures

3.1	Methodology	32
4.1	Screenshot of the Collected Dataset	37
4.2	Screenshot of the ILPD Dataset	37
4.3	Screenshot of the Combined Dataset	38
5.1	Precision, Recall & F1-score Comparison of Models on the Test Dataset	42
5.2	Precision, Recall & F1-score Comparison of Models on the Validation Dataset	44
5.3	Accuracy Comparison across Models on the Test Dataset	45
5.4	Accuracy Comparison across Models on the Validation Dataset	45
5.5	Comparison of Confusion Matrices across Models	48
5.6	Confusion Matrices of the Stacked Ensemble Model	49
5.7	SHAP Beeswarm Plot of the Stacked Ensemble Model	51
5.8	SHAP Feature Importance Plot of the Stacked Ensemble Model	51
5.9	SHAP Waterfall Plot of the Stacked Ensemble Model	52
5.10	SHAP Dependence Plot of the Stacked Ensemble Model	52
6.1	Workflow of the Website	55
6.2	LiverGuard Home Page Interface	57
6.3	Prediction Input Form	58
6.4	Prediction Input Form with Data and Model Request Preview	59
6.5	Prediction Result Page with Risk Output	60
6.6	Model Performance Page – Evaluation Metrics and Confusion Matrix	61
6.7	Model Performance Page – Ensemble Signals and System Details	62
6.8	About Page – Project Overview and Explainability Components	63
6.9	About Page – Objectives, Workflow, and Model Details	64
6.10	Team Page - Project Members and Supervisor	65

List of Tables

2.1	Machine Learning–Based NAFLD Prediction Literature Analysis . . .	14
2.2	Matrix Decomposition-Based Literature Analysis	28
2.3	Overview of NAFLD Datasets	30
5.1	Class-Wise Test Performance Metrics Across Classification Models .	46
5.2	SHAP Explainability Report of the Stacked Ensemble Model	53

ABSTRACT

Non-Alcoholic Fatty Liver Disease (NAFLD) is an emerging health issue across the world especially in less developed and middle-income nations like Bangladesh where early diagnosis is essential to avoid severe liver-related complications but the current screening procedures are invasive, costly, and not suitable to be applied at large scale. This paper proposes an interpretable machine learning-based model of NAFLD risk prediction on the basis of a unified set of heterogeneous publicly available data and locally-gathered real-world clinical data reflecting the Bangladeshi population. The combined dataset had a significant amount of missing data as a result of dissimilar feature presence that was filled in with low-rank matrix decomposition with the help of SoftImpute. SMOTE oversampling has been used to reduce the imbalance in the classes after the completion of the matrices. On the final balanced dataset, the Logistic Regression, Random Forest and XGBoost models were trained as well as a stacking ensemble model. Experimental findings confirm that Logistic Regression had a test accuracy of 97.99%, whereas Random Forest, XGBoost and the stacked ensemble had test accuracy of 99.28%, 99.31% and 99.40% with the ensemble giving perfect class discrimination with a ROC AUC of 100% and almost perfect class-wise precision and recall. The SHAP-based explainability was added in order to offer both overall feature significance and patient-level clarifications.

Keywords: Non-Alcoholic Fatty Liver Disease (NAFLD), Explainable AI, Machine Learning, Clinical Decision Support, Good health and Well-being

Chapter 1

Introduction

Non-alcoholic fatty liver disease (NAFLD) has become one of the most prevalent chronic liver diseases in the world with close relation to metabolic dysfunction. Which includes obesity, diabetes type 2, insulin resistance, hypertension, and dyslipidemia [1]. According to epidemiological research, around 25-30 percent of adults are infected around the globe with even greater success in the case of people having metabolic syndrome [3]. In South Asia, such as Bangladesh, the risk of NAFLD is increasing due to the rapid urbanization and change of lifestyle, which highlights the necessity to develop scalable and timely risk assessment tools [5, 18]. Liver biopsy is still the clinical gold standard, but its invasiveness, high cost, and the risks of the procedure do not permit extensive screening [7, 11]. Non-invasive methods including ultrasound, controlled attenuation parameter (CAP) and MRI-based methods provide only partial solutions but lack specificity to early disease stages [10], [29]. As a result, machine learning (ML)-based predictive models that rely on clinical and laboratory data gathered routinely have become promising and cost-effective methods of NAFLD risk detection [6], [9]. Classical ML methods like logistic regression, support machine learning and random forests have shown moderate predictive performances [13]. Experts like ensemble learning and hybrid systems like bagging, boosting, and stacking further enhanced quality and strength [4, 14]. In addition, explainable AI (XAI) methods, in particular SHAP-based interpretations have been implemented in order to increase transparency and clinical interpretability, which enables the identification of the most important risk factors on an individual and population scale [8, 16, 24, 28]. One of the weaknesses with current research is the use of a single-center or single-cohort dataset with a homogeneous feature space, which prevents generalizability and exposure to

heterogeneous clinical profiles [2, 3, 19]. Public datasets are diverse in size, demographics and available features, and frequently result in incomplete or fragmented datasets. Though there are studies that compare models between multiple datasets, they usually do so by performing independent training and tests without the use of complementary information by using integrated feature-level fusion [17, 19].

The combination of heterogeneous datasets creates structural missingness, in which some features are not available in selected datasets, thus forming sparse feature matrices. Common imputation techniques, including mean, median, or KNN imputation, are not able to reveal multifaceted latent correlations, and this can bias clinically significant patterns [30, 31]. It is believed that advanced methods such as matrix decomposition and completion can be used successfully in other aspects of biomedical science to recover lost values using low-rank structures and latent correlations, even in high-dimensional, sparse, or nonlinear settings [32, 33].

Although they have succeeded in genomics, imaging, and drug discovery [30, 31, 32, 33], they have hardly been used in predicting NAFLD risks. Specifically, there is no research so far that fuses multi-dataset features, completes missing values based on matrix decomposition, ensemble ML, and can be explained in an interpretable manner into one unified structure. Current ensemble models generally act on a single dataset, and explainability has generally been added afterwards without considering the heterogeneity of feature availability [4, 16, 17].

1.1 Problem statement

Non-Alcoholic Fatty Liver Disease (NAFLD) has risen as one of the world's most widespread chronic liver diseases. This is strongly associated with metabolic conditions, for example, obesity, diabetes, and hypertension. In countries like Bangladesh, the prevalence is increasing due to lifestyle changes and urbanization.

Current diagnostic methods, such as liver biopsy, are invasive, expensive, and not suitable for large-scale screening. Non-invasive techniques exist, but they often lack accuracy in early-stage detection. Machine learning-based approaches have shown promise, but most existing systems rely on single datasets, limiting their generalization ability.

Additionally, combining multiple datasets introduces missing values due to feature differences, and traditional imputation methods fail to capture complex

relationships in clinical data. There is also a lack of integrated systems that combines dataset fusion, advanced imputation, ensemble learning, and explainability in one framework.

Therefore, the problem is to design a scalable, accurate, and interpretable NAFLD prediction system that can work with heterogeneous datasets and support real-world clinical decision-making.

1.2 Aims, objectives and Motivation

1.2.1 Aims

The main aim of this project is to develop an explainable machine learning-based system for accurate and scalable prediction of NAFLD risk using both public and real-world clinical data.

1.2.2 Objective

The specific objectives of this project are:

- To combine multiple heterogeneous public datasets with a local clinical dataset into a unified feature space
- To handle missing data using matrix decomposition (SoftImpute)
- To balance the dataset using SMOTE for better model performance
- To design a feature-level dataset fusion strategy that aligns and merges multiple public NAFLD datasets through feature harmonization, label standardization, and zero-padding, enabling unified model training.
- To train and compare multiple machine learning models such as, Logistic Regression, Random Forest, and XGBoost
- To develop a stacking ensemble model for improved prediction accuracy
- To assess model performance using standard classification metrics across training, validation, and test sets to ensure robustness and prevent overfitting.

- To apply SHAP-based explainability to the best-performing models in order to derive global feature importance and local, patient-level explanations.
- To build a system that supports early detection and clinical decision-making

1.2.3 Motivation

NAFLD is increasing rapidly, especially in developing countries, but early diagnosis remains difficult due to costly and invasive methods. Many patients remain undiagnosed until severe complications occur. Existing machine learning solutions are often limited by small datasets, lack of generalization, and poor interpretability. This motivated us to develop a system that not only improves prediction accuracy but also provides clear explanations for clinical use. By combining multiple datasets, handling missing values effectively, and applying explainable AI, this project aims to create a practical and reliable solution that can support large-scale screening and assist healthcare professionals in early risk assessment.

1.3 Project Specification

The specifications of this project are outlined below.

1.3.1 Input Features

The model uses routinely available clinical and laboratory features, including:

- Demographic information (Age)
- Height, Weight and BMI
- Fasting Blood sugar
- Liver enzymes (ALT, AST)
- Lipid profile (Total Cholesterol, HDL, LDL, Triglycerides)
- Diabetes status etc.

1.3.2 Prediction Targets

- NAFLD Risk (Yes/No classification)

1.3.3 Performance Goals

- To achieve high classification accuracy (99%) on test data
- To ensure high Recall for NAFLD cases to minimize false negatives
- To maintain balanced performance
- To achieve stable performance across multiple heterogeneous datasets
- To provide global and patient-level explanations using SHAP

1.4 Expected Outcomes

The expected outcomes of this project are as follows:

1.4.1 Model Training Outcomes

- To validate machine learning models for NAFLD risk prediction using clinical features
- To identify the most reliable and clinically feasible model

1.4.2 Explainability Outcomes

- Integrating SHAP-based explainability to ensure transparency
- Generation of patient-level explainability reports, including: Risk level (High / Moderate / Low), probability score and key contributing features

1.4.3 System Outcomes

- Development of a web-based clinical decision support system
- A user-friendly interface where clinicians can input patient data and receive: NAFLD risk prediction and a visual explanations of contributing risk factors

Chapter 2

Literature Review

2.1 NAFLD: Epidemiology, Risk Factors, and Diagnostic Challenges

The epidemiologic research and ML-based screening solutions tend to suggest that NAFLD has a significant proportion of an adult population in the global population, and prevalence rates are usually approximately 25-35 per cent and even greater among some population groups with obesity or diabetes [1, 3, 10]. Asian population studies, such as the South Asian countries, reveal that NAFLD is very common in non-Western nations, which underlines the disease burden on the global scale[5, 18]. In classical NAFLD ML articles, the risk factors most frequently identified are the presence of increased BMI and waist circumference, the elements of metabolic syndrome, the presence of high fasting glucose or HbA1c, dyslipidemia, and high liver enzymes, such as ALT, AST, and GGT [1, 3, 6, 9]. Sex, age and hypertension are also other significant predictors as noted by a few models[4, 6, 19]. In the studies, a diagnosis of NAFLD is typically determined with the aid of abdominal ultrasound, transient elastography with CAP, or MRI-PDFF, and only a minor proportion of the studies relies on biopsy as ground truth, making it heterogeneous and subject to misclassification [1, 2, 7, 10]. These diagnostic limitations encourage the development of effective, non-invasive prediction models as an addition or prioritization of imaging and biopsy.

2.2 Interpretability in NAFLD AI and ML Research

2.2.1 Background and Importance

The increasing applications of ML and AI to predict NAFLD have also brought issue of model transparency since most of the high-performing models are non-interpretable and hard to use [8, 14, 17]. The scientific reviews on AI in NAFLD and liver fibrosis underline that although the reported values of the use of the models are frequently high, not much attention is paid to the way clinicians interpret and apply these models in practices [7, 11].

The interpretability is also especially significant in NAFLD as the predictions can impact the decision-making regarding further testing, lifestyle changes, or the change of medication. Explainable ML techniques may help to explain which risk factors contribute to a prediction of a given patient and may therefore be more readily verifiable by the clinician when compared to the pathophysiologic knowledge and patient history [8, 12, 16]. This view promotes the inclusion of interpretability as an initial component in the design of models instead of an appendage.

2.2.2 Key Interpretability Approaches in Reviewed Studies

The more recent NAFLD ML works are starting to apply Shapley Additive explanations (SHAP) to measure feature effects both globally and individually, particularly in tree-based models like gradient boosting or XGBoost [8, 14, 16]. As an example, SHAP is used to show that BMI, liver enzymes, and lipid measures, and measurements of visceral adiposity are strong contributors to the model prediction in multi-cohort NAFLD and fibrosis prediction studies [14, 22, 23].

Other tools used to explain scenarios have been used with SHAP. A study on NAFLD web-applications involved the application of neural networks followed by the visualization of the effect of age, GGT, and abdominal measurements on the risk scores of the individual users using both SHAP and LIME [12]. Ensemble models which couple many classifiers also provide feature-importance rankings and decision-path exploration to shed light on why ensemble predictions are superior to single algorithms in particular cohorts [4, 17, 21]. Attention maps and associated visualization systems have been applied in imaging-based NAFLD detection to

indicate image regions that point to steatosis, but these experiments continue to tend to supplement image explanations with table-based feature importance [7, 10].

2.2.3 Trends and Limitations

One such trend in the literature is the progressive increase in the literature in which some form of explanation (usually global feature importance or SHAP summary plots) is provided in addition to performance metrics [8, 12, 14, 16]. Nevertheless, interpretability is not the most common final goal and most of the studies incorporate only a limited number of visualizations in the findings without providing a formal structure of how explanations are to be read or applied by clinicians [7, 11, 24].

Several limitations remain. Post-hoc models like SHAP tend to be used on complex models that were not intended to be interpretable, and thus making explanations susceptible to data changes [16, 19]. Furthermore, very little NAFLD literature formally evaluates with clinicians the concepts of whether explanations do in fact enhance understanding, trust, or decision-making and some studies indicate that explanations are not necessarily enough to ensure usability [24]. The other forms of explanation like counterfactuals, contrastive arguments or the type of explanations which are sensitive to uncertainty are also not thoroughly investigated in the NAFLD field.

2.3 Machine Learning in NAFLD Prediction

2.3.1 Overview of ML Applications

ML has been used in a broad range of NAFLD related tasks, most frequently, binary classification of NAFLD vs non NAFLD with the help of clinical and laboratory data [1, 3, 6, 9]. It can also be used to predict more advanced fibrosis or NASH/MASH, predict mortality risk and also identify high-risk subgroups including lean NAFLD [12, 13, 16, 23]. Various studies also use ML together with national survey data or multi-cohort samples to investigate NAFLD-related results on the population level [2, 4, 10, 19]. The majority of classical NAFLD models are trained using tabular data, and they are often reported to perform well, with the undefined performance measures being the range of 0.85-0.95 in binary NAFLD prediction [2, 4, 14, 21]. Ultrasound images, facial photographs, and multimodal combinations of clinical

and imaging features have also been subjected to deep learning which in some cases can have very high sensitivity and specificity but can often lack interpretability and data needs [5, 7, 20].

2.3.2 Commonly Used Algorithms

The literature shows a consistent preference for certain algorithm families:

- Tree-based ensembles: The approaches, which include gradient boosting, XGBoost, and random forests are some of the most common and typically perform the best in most NAFLD datasets [2, 3, 4, 14].
- Linear models: Interpretable baselines Generalized linear models and logistic regression are popular, and even competitive, in mortality studies or risk-score based studies [3, 6, 13, 23].
- Support vector machines and k-nearest neighbors: These are both typically featured in comparative studies, and may also do well in some feature spaces, or with some high-quality preprocessing [3, 6, 9, 15].

More generally, stacked or voting ensembles that are composed of a collection of base learners are also discussed and can often achieve state-of-the-art performance on particular cohorts, although the added interpretability tooling is widely necessary to ensure the results are clinically comprehensible [4, 17, 21].

2.3.3 Performance Metrics in Literature

The most widespread reported performance measure of NAFLD prediction models is AUROC, which is commonly reported to take values above 0.85 with models based on tree-based ensembles or a well-tuned logistic regression [2, 4, 14, 21]. Accuracy, sensitivity, specificity, precision, and F1-score have also been reported in numerous studies, especially comparing between multiple algorithms on one dataset [3, 6, 9, 15]. Others include calibration plots or decision-curve analyses, which are used to indicate whether a model can give useful information to the clinical decision maker instead of ranking of risk [1, 13, 25]. The imaging-based AI systems of NAFLD and other liver diseases demonstrate extremely high pooled sensitivity and specificity when operating in controlled settings, although reproducibility and performance in the real-world context are possibly worse [7, 11].

2.3.4 Challenges and Limitations

Some of the most significant challenges and limitations mentioned in NAFLD literature on ML are:

- Limited generalizability: Most models are constructed, often using a single-center or country-specific cohort, e.g. health-check populations of East Asia or individual hospitals, which may lower their generalizability to other populations or clinical environments [1, 5, 19, 25].
- Lack of external validation: It is not externally validated on independent, external cohorts, but only a small proportion of the models are externally validated by cross-validation of internal validation, or single split-validity [2, 3, 14].
- Variable reference standards: The differences in the definition of NAFLD (based on ultrasound, elastography, MRI-PDFF, or non-invasive index) introduce a heterogeneity of the labels, and it becomes hard to compare models directly [1, 2, 7, 10].
- Handling of imbalance and missing data: NAFLD datasets often show class imbalance and missing values; methods such as SMOTE or related oversampling strategies are employed but not always evaluated thoroughly [4, 6, 15, 17].
- Limited longitudinal modeling: The majority of studies use cross-sectional data, and they do not predict the progression of the disease, yet longitudinal prediction of either fibrosis or mortality is demonstrated to be a clinically relevant aspect [11, 13, 16, 23].

These obstacles indicate that more robust modeling approaches, improved reporting, and broad-based analysis of various data should be conducted.

2.3.5 Identified Gaps

The reviewed studies have a number of gaps that can be traced:

- Systematic interpretability frameworks SHAP or feature importance plots have been highly utilized in various contexts, yet they have not been systematically used to improve or evaluate the implication of explanations on clinician behavior [8, 12, 24].
- Cross cohort strength and cross-population consistency of explanation: Not many studies also determine the effect of features on or consistency of explanation across population or domain shift, which is needed for trustful deployment [2, 14, 19].
- Outcome-based explainable models: Despite the fact that most of the models are aimed to identify NAFLD, fewer of them generate explainable models that are specifically on fibrosis progression, NASH/MASH or mortality but are essential clinical outcomes [11, 13, 16, 23].

The proposed project aims to fill the above gaps by creating a model of NAFLD that can be interpreted, analyzing the tendencies of the explanation.

2.4 Handling Missing Data using Matrix Decomposition Techniques

Handling missing data is a critical challenge in NAFLD prediction, especially when integrating multiple heterogeneous datasets. Traditional imputation methods such as mean, median, or KNN imputation are commonly used. However, they often fail to capture complex nonlinear relationships between clinical features. This can lead to information loss and biased predictions [24, 25, 29].

Recent studies have shown that matrix completion and low-rank decomposition techniques provide more effective solutions for handling missing data in biomedical applications. For instance, Fan et al. [30] proposed a deep matrix completion method that integrates radiomic and histological data to recover missing values. And demonstrated improved performance by learning nonlinear latent feature representations. Similarly, Huang et al. [31] introduced a low-rank matrix completion approach for single-cell RNA sequencing data. Which preserves cell heterogeneity while improving imputation accuracy.

Du et al. [32] further extended this concept by proposing a meta-learning-based inductive logistic matrix completion method. This method improves prediction performance in sparse datasets by effectively modeling feature interactions. In addition, nonlinear matrix completion techniques have been developed to capture complex dependencies between variables using multi-layer transformations. This showed strong performance in various high-dimensional applications [33].

Despite these advancements, the application of matrix decomposition techniques in NAFLD prediction remains limited. Most existing studies still rely on simpler imputation strategies that may not fully utilize the underlying structure of heterogeneous clinical data.

To address this limitation, the present study adopts a matrix decomposition-based imputation approach (SoftImpute) to handle structural missingness created by dataset fusion. This method leverages low-rank approximation to estimate missing values while preserving latent relationships among features. By integrating this approach into the machine learning pipeline, the proposed framework improves data quality, enhances model performance, and enables more reliable NAFLD risk prediction.

2.5 Dataset Summary in Classical NAFLD Studies

Studies of classical NAFLD ML use diverse types of datasets. A number of high-impact studies apply large population based surveys or national health examination data which offer standardized anthropometric and laboratory variables and in a number of studies elastography based measurements of hepatic fat and fibrosis [2, 4, 10, 19]. Such datasets typically have thousands of participants and can be used in the creation of risk-prediction models of general applicability but can be limited in their applicability to regions or time. The other category of many studies is based on hospital-based or health-check cohorts, including those of single or multi-centers, which were obtained at outpatient clinics or check-up programs [1, 3, 5, 25]. These studies have sample sizes of hundreds up to hundreds of thousands, and are usually well-characterized by biochemical and anthropometric measures, but they may also be subject to selection bias and no longer-term follow-up [6, 15, 21]. Fewer articles incorporate specialized imaging or multimodal data, e.g. MRI-PDFF, ultrasound-advanced or a combination of facial images and clinical data, which can be used to

provide accurate labeling of steatosis and fibrosis but are costly and not as widely accessible [7, 20]. There are also models that are trained on published ML datasets such as ILPD or other liver-disorder datasets that can be found in repositories and competitions making it easy to compare algorithms but not necessarily to clinical NAFLD definitions or target populations [5, 9, 17].

2.6 Machine Learning–Based NAFLD Prediction Literature Analysis

Table 2.1: Machine Learning–Based NAFLD Prediction Literature Analysis

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
1	Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study	To compare and develop ML models for early identification of NAFLD.	XGBoost model achieved the highest scores across multiple evaluation metrics, including (AUROC), (AUPRC), accuracy, positive predictive value, F1 score, (MCC), and specificity.	Developed multiple ML models; XGBoost performed best. Achieved AUROC = 0.873, Accuracy = 0.795. Identified BMI as the most important predictor.	Retrospective, cross-sectional design (no causal inference). Chinese-only dataset, limiting generalizability. Diagnosis based on ultrasound, not biopsy. No external validation across cohorts.	Retrospective Chinese clinical dataset (15,315 subjects; split into 10,373 training and 4,942 testing)
2	Automated machine learning models for nonalcoholic fatty liver disease assessed by controlled attenuation parameter from the NHANES 2017-2020	To evaluate automated machine learning (AutoML) for NAFLD detection in population-based cross-sectional databases.	GBM, DRF, XRT, GLM, XGBoost, Deep Learning.	XGBoost model achieved AUC 0.859, Accuracy 0.795, Sensitivity 0.773, Specificity 0.802. Developed R Shiny web app “Shiny NAFLD.”	Same NHANES database used for both training and validation, limiting generalizability. Lack of external validation with datasets from diverse demographics and settings. Uncertainty around CAP cutoff values for steatosis diagnosis remains unresolved.	NHANES (2017–2020), 4,177 participants

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
3	Development and validation of machine learning models for nonalcoholic fatty liver disease	To develop and validate novel machine learning models for predicting nonalcoholic fatty liver disease (NAFLD) by combining various risk factors.	Logistic Regression, Random Forest, XGBoost, GBM, SVM.	XGBoost model performed best (AUC = 0.938 & 95% CI: 0.870-0.950) Reliable reference for early identification of high-risk NAFLD patients in clinical practice.	The sample size for the testing set (131 participants) was relatively small compared to the training set (578 participants) Limited external validation	Training Set: 578 participants (329 with NAFLD, 249 without NAFLD), with abdominal ultrasound examinations. Testing Set: 131 participants (96 with NAFLD, 35 without NAFLD), with magnetic resonance imaging (MRI) for external validation.
4	Precision Non-Alcoholic Fatty Liver Disease (NAFLD) Diagnosis: Leveraging Ensemble Machine Learning and Gender Insights for Cost-Effective Detection	To develop a cost-effective tool for early diagnosis of (NAFLD) using affordable and accessible laboratory test data, aiming for better management of the condition.	Random Forest, KNN, Logistic Regression, SVM, XGBoost, Decision Tree	Ensemble model achieved Accuracy 0.99, AUC 1.00. Outperformed other ML techniques and the conventional hepatic steatosis index (HSI).	Use of restricted datasets limits study scope. Absence of clinical data reduces clinical relevance. No clinical trial conducted to validate findings.	NHANES dataset (2017–2020)
5	Detection of Nonalcoholic Fatty Liver Disease Using Deep Learning Algorithms	To examine fatty liver signs and disorders to help diagnose diabetes early using deep learning algorithms for the detection of Nonalcoholic Fatty Liver Disease (NAFLD).	CNN, LSTM, RNN, MLP	Deep learning achieved 71% accuracy in a highly categorized environment for predicting fatty liver disease.	Data collected from a single medical center. Lack of multi-center datasets and external validation reduces reliability. Small sample size (584 patients)	ILPD Dataset from the UCI Machine Learning Repository, consisting of 583 data points, each with 11 features related to liver health.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
6	Machine Learning Algorithms Based Non Alcoholic Fatty Liver Disease Prediction	To identify potential factors causing NAFLD using various Machine Learning algorithms and to predict NAFLD at early stages, which is crucial for timely treatment and prevention of further health complications.	Decision Tree, SVM, Random Forest, Logistic Regression.	The study uses accuracy as the parameter for performance analysis evaluation. Random Forest model accurately predicts non-alcoholic fatty liver disease in patients.	Limited and cleaned dataset No external validation or multicenter data, limiting broader applicability. Excludes imaging data like ultrasound, which may provide essential diagnostic information.	Custom dataset: 605 patients with 27 features related to NAFLD, reduced to 451 entries after preprocessing.
7	Artificial Intelligence for Detecting and Quantifying Fatty Liver in Ultrasound Images: A Systematic Review	To explore the effectiveness and performance of various Artificial Intelligence (AI) methods applied to ultrasound images for diagnosing and quantifying (NAFLD)	Involves the use of multiple Artificial Intelligence methods combined with ultrasound imaging for diagnosing NAFLD.	AI significantly improved the diagnosis of NAFLD, Non-Alcoholic Steatohepatitis (NASH), and liver fibrosis AI-supported systems showed potential performance increases in detecting and quantifying steatosis, NASH, and liver fibrosis in NAFLD patients	Need for further validation Prospective studies required before implementing AI-supported systems in real-world scenarios	49 reviewed studies (various clinical cohorts).

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
8	Advancing non-alcoholic fatty liver disease prediction: a comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation	To develop and validate advanced machine learning models for predicting NAFLD, integrating SHAP interpretability to enhance understanding of predictive factors.	7 ML models; LightGBM performed best.	SHAP method was used to enhance model interpretability. The light gradient boosting machine model achieved the highest performance. AUC = 0.90 (internal), 0.81 (external). Accuracy = 87%, Sensitivity = 92.9%, and F1 score = 0.92.	Longitudinal data and lifestyle factors should be included to further refine risk assessments.	Dryad + NHANES (2017–2020). 14,913 participants.
9	Early Identification of Risk Factors in Non-Alcoholic Fatty Liver Disease (NAFLD) Using Machine Learning	To explore the use of Machine Learning Algorithms (MLAs) for identifying risk factors and predicting the development of NAFLD.	The research utilized 9 different ML Algorithms across four datasets and ensemble modeling techniques such as bagging and boosting was employed.	ML Algorithms were highly effective in predicting NAFLD risk factors, both when used individually and as part of ensemble modeling techniques. Identified the top attributes associated with NAFLD risk across all datasets	Small sample sizes in some datasets Lack of gold-standard diagnostics (no liver biopsy or FibroScan in several datasets; reliance on blood markers only). Dataset bias and overfitting risks. -Incomplete clinical variables (e.g., missing education, family history, dietary standardization in certain datasets).	Liver Disorders Data Set, HCC Survival Dataset, ILPD Dataset (Indian Liver Patient Dataset), Cirrhosis Prediction Dataset

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
10	Non-invasive diagnosis of non-alcoholic fatty liver disease: Current status and future perspective	To review and discuss non-invasive diagnostic methods for NAFLD and non-alcoholic steatohepatitis (NASH), focusing on recent developments in biomarkers and imaging techniques for improved diagnosis, assessment, and monitoring of the disease.	1. Omics techniques for serological assays 2. Imaging techniques, including Three-dimensional magnetic resonance imaging (3D-MRE) 3. Combination tests like FM-fibro LSM index 4. Machine learning algorithms (MLA) for liver fibrosis assessment.	Highlighted promising non-invasive tests (3D-MRE, FM-fibro index).	Need to be validated through larger cohort studies to confirm their accuracy and reliability.	For NAFLD Diagnosis: 20-carboxy arachidonic acid (20-COOH AA) 13,14-dihydro-15-keto prostaglandin D2 (dhk PGD2) For Liver Fibrosis Assessment: Three-dimensional magnetic resonance imaging (3D-MRE) FM-fibro index combined with Liver stiffness measurement (FM-fibro LSM index)
11	Application of artificial intelligence in non-alcoholic fatty liver disease and liver fibrosis: a systematic review and meta-analysis	To estimate the performance of AI-assisted systems for diagnosing NAFLD, Non-alcoholic steatohepatitis (NASH), and liver fibrosis, and to determine this performance through a systematic review and meta-analysis.	Neural networks (ANN, CNN) and non-neural network models (RT, rule extraction, Lasso regression). ELM, SVM, NB, RF, BN, LR, KNN, AdaBoost, HNB, Bagging, AODE, SGD, XGBoost, MLP, and LSTM.	AI-assisted ultrasonography for NAFLD showed excellent performance (sensitivity 0.97, specificity 0.98, AUC 0.98), outperforming clinical datasets. For NASH, clinical AI models achieved sensitivity 0.80, specificity 0.69. For liver fibrosis (F1-F4), sensitivity was 0.99-1.00, specificity 0.76-1.00.	Uncertain bias in some studies, mixed alcoholic/non-alcoholic fatty liver disease - Lacked steatosis distribution details. Limited non-neural model validation	25 studies on AI in diagnosing NAFLD, NASH, and liver fibrosis, focusing on ultrasonography-based AI for NAFLD and AI-assisted clinical datasets for all conditions.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
12	Development and validation of a neural network for NAFLD diagnosis	To develop and validate a simple Neural Network (NN)-based web application capable of predicting NAFLD and to reduce unnecessary ultrasounds.	Neural Network (MLPClassifier), Python along with "flask" for the web application interface. HTML, CSS, JavaScript, (t-SNE), LIME and SHAP	NN achieved 77 percent accuracy, 0.82 of AUC -Specificity 1.00 and sensitivity of 0.73 for healthy subjects. Highlighted AVI, Age, and GGT as the most significant features in NAFLD diagnosis.	Inaccessibility of certain imaging or laboratory parameters for diagnosis High cost and impracticality of existing imaging technologies Misclassification of some mild NAFLD cases as healthy.	2,970 subjects (train/test), 2,301 subjects (validation).
13	Machine learning-based mortality prediction models for non-alcoholic fatty liver disease in the general United States population	To find a generalizable and valuable model for early mortality prediction in ultrasound-proven NAFLD patients and compare their performance with conventional non-invasive scoring systems.	Logistic Regression, K-nearest neighbors (KNN), Gradient-boosted decision tree (XGBoost), Random Forest, and Decision Tree.	The Logistic Regression (LR) model achieved an AUC of 0.888, outperforming other ML models -All ML models surpassed traditional non-invasive scores. The Decision Tree model, using age, systolic pressure, and glycated hemoglobin, was found practical Age was the most critical predictor across all models.	Reliance on ultrasound for NAFLD diagnosis -Missing data within the NHANES III dataset. Lack of a comparable external dataset for validation.	NHANES-III (3,233 patients, 25.3 years follow-up).

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
14	Development and Validation of Machine Learning-Based Marker for Early Detection and Prognosis Stratification of Nonalcoholic Fatty Liver Disease	To develop a multimodal stacked ML model that incorporates genetic and clinical data to synthesize an in-silico quantitative marker for Nonalcoholic Fatty Liver Disease (NAFLD) (ISNLD).	Logistic Regression, LGBM, XGBoost, SVM, Random Forest, ADB, KNN and Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection.	Multimodal model achieved strong performance: AUCs of 0.843 (training), 0.840 (internal test), and 0.872 (external test). Showed good prediction of SeLD (AUC 0.787 in high-risk NAFLD) and potential to assess extrahepatic metabolic risks.	Relies on clinical features only. Limited continuous-scale quantification.	1. Primary Dataset: 2,970 subjects (2,869 used for training/testing via a train-test-split approach). 2. Secondary Dataset: 2,301 subjects, with 100 randomly selected to test the web application.
15	Classification of NAFLD Using ML Algorithms	To improve the predictive accuracy of three NAFLD stages: early (F0 and F1), intermediate (F2), and severe (F3 and F4), using ML techniques & to predict the stage of NAFLD in biopsy-proven patients using ML algorithms, providing a non-invasive solution to health practitioners.	Random Forest (RF), Extra Tree Classifier (ETC), Multi-layer Perceptron (MLP), Logistic Regression, Histogram Gradient Boosting (HBG), and Gradient Boosting (GB).	Random Forest achieved 68.75% accuracy, an F-1 score of 68.75%, weighted sensitivity of 68.75%, positive predictive value of 68.85%, and negative predictive value of 84.38%.	The training data utilized was relatively limited because it was biopsy-proven NAFLD, potentially impacting the generalizability of the findings. There was also a class imbalance in the dataset, which was addressed using oversampling techniques like SMOTE and data imputation (Multivariate imputer) for missing values.	Biopsy-proven NAFLD dataset.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
16	An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis	The study aimed to develop an explainable ML model for predicting high-risk metabolic dysfunction-associated steatohepatitis (MASH) and compare its performance with established biomarkers.	XGBoost model for detecting high-risk MASH SHAP method for model interpretability.	The XGBoost model achieved high performance metrics, including sensitivity (0.82), specificity (0.91), accuracy (0.90), and AUC (0.95), outperforming established biomarkers like FIB-4, APRI, BARD, and MASLD fibrosis scores in predicting high-risk MASH.	The study lacked liver biopsy data for gold-standard comparison. The integration of the optimized model into clinical practice and electronic medical records has not yet been evaluated. Lack of external cohort validation	NHANES dataset (2017–March 2020): 5,281 adults with valid elastography measurements.
17	Stacking Ensemble Learning Approach for Non-Alcoholic Fatty Liver Disease Identification: Leveraging Explainable Machine Learning for Enhanced Prediction Models	To develop an efficient detection method for NAFLD using a stacked ensemble approach that integrates multiple machine learning models and provides model interpretability.	Stacking ensemble using base classifiers–XGBoost, CatBoost, Decision Tree, and AdaBoost & using SHAP for feature importance.	The stacked ensemble model achieved a classification accuracy of 95.9%, outperforming individual models and demonstrating the robustness and reliability of the ensemble approach for NAFLD detection.	Reliance on a specific dataset Lack of exploration on diverse patient cohorts and additional clinical settings for widespread clinical application	NAFLD dataset from Kaggle. It includes adult patient data from 1997–2014 with demographic and clinical risk attributes.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
18	Explainable Artificial Intelligence for Diagnosis and Staging of Liver Cirrhosis Using Stacked Ensemble and Multi-Task Learning	To develop a deep learning framework that combines stacked ensemble and multi-task learning to diagnose and stage liver cirrhosis using MRI images with explainability.	Base models: VGG16, MobileNet, DenseNet121; meta-classifier: XGBoost. Implemented using TensorFlow 2.16.2; Grad-CAM for explainability.	Achieved a mean accuracy of 96.92% in binary classification and 96.71% in staging (AUC: 99.81%), outperforming single-model baselines and demonstrating superior performance and robustness across cross-validation folds.	Imbalance in the dataset (fewer advanced-stage cases). Cross-validation may suffer from overfitting due to lack of patient-wise separation. Imaging protocol variations and lack of radiologist involvement in validation.	CirrMRI600 + T2-weighted liver MRI images labeled with cirrhosis stages, used for classification and staging evaluation.
19	Application of Artificial Intelligence in Support of NAFLD Diagnosis	To construct a modular AI-based diagnostic system for NAFLD that integrates machine learning models, interpretable outputs, and data pipeline automation.	Random Forest, XGBoost, SVM, Logistic Regression, MLP, AdaBoost, KNN, Decision Tree & SHAP for explainability	The ensemble model combining several base models achieved an accuracy of 89% and an ROC AUC of 89% on the test set, demonstrating high diagnostic performance through collaborative decision-making of complementary individual models.	Based on a single pediatric dataset from Warsaw, limiting demographic diversity. External clinical variables such as diet and lifestyle were not incorporated. Prospective clinical validation is yet to be performed.	Clinical dataset from Children's Health Center, Warsaw (2015–2023); includes blood markers, anthropometrics, and ultrasound data.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
20	Multi-modal Learning based Prediction for Disease	To build a deep learning framework combining facial imaging and clinical metadata for early NAFLD detection, offering a low-cost and non-invasive alternative to liver biopsy.	3dMDface system for image capture; CNNs for facial feature extraction; MLP for clinical features; SHAP for interpretability; compared with Random Forest and metadata-only baselines.	The DeepFLD model combining facial images and metadata achieved the best NAFLD prediction with an AUC of 0.923, accuracy of 87.5%, specificity of 82.5%, and positive predictive value of 83.9%.	Generalizability beyond the North Chinese population remains uncertain. Imaging conditions (lighting, exposure) varied across collection years. Diagnosis based on ultrasound (not biopsy), introducing uncertainty.	FLDDData: facial photos & metadata from NAFLD patients in China collected over 5 years using a high-fidelity 3D imaging system.
21	Multimodal Learning-based Prediction for Nonalcoholic Fatty Liver Disease	To enhance NAFLD prediction accuracy by combining 3D facial images with multidimensional clinical metadata through deep multimodal fusion.	3dMDface system; CNNs for facial data; MLP and Vision Transformer (TransFG) for feature fusion; SHAP for interpretation.	Fusion model achieved AUC = 0.923, Accuracy = 87.5%, Specificity = 82.5%, and PPV = 84.5%. Image-only models (AUC ~0.863) showed facial morphology carried relevant disease signals. -Model remained robust across multiple years of data collection, proving its generalizability.	Limited to a specific geographical cohort. -SHAP explanations improved transparency, but full interpretability remains challenging in deep models. Relied on ultrasound-based labeling rather than biopsy-confirmed NAFLD.	Dataset: FLD-Data

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
22	Machine learning prediction of metabolic dysfunction-associated fatty liver disease risk in American adults using body composition: explainable analysis based on SHapley Additive exPlanations	To predict MAFLD risk using body composition variables and explainable ML models, aiming to identify key adipose tissue indicators contributing to liver disease.	ML models: Decision Tree, SVM, GLM, GBM, Random Forest, XGBoost; Feature selection: Boruta; SHAP for interpretability; R environment with caret and DALEX.	GBM showed best performance: AUC = 0.879 (validation), Sensitivity 0.78, Specificity 0.83. SHAP identified VAT, BMI, and subcutaneous fat as key predictors. Ensemble models outperformed linear ones, highlighting the complexity of adiposity-related risk	Diagnosis relied on FibroScan, not biopsy. Cross-sectional design restricts causal inference. Model tested only on NHANES data without external validation.	Survey dataset (NHANES / US-based survey) + UCI references
23	Predicting the risk of lean non-alcoholic fatty liver disease based on interpretable machine models in a Chinese T2DM population	To create ML models capable of predicting NAFLD among lean type 2 diabetes mellitus (T2DM) patients—an underexplored subgroup at risk.	ML algorithms: Random Forest, Logistic Regression, LDA, SVM, XGBoost, Naive Bayes; Feature selection: Boruta + LASSO; SHAP for interpretation; built in Python & R.	Among 1,553 T2DM patients (20.3% with lean NAFLD), the Random Forest model performed best with an AUC of 0.739 (training) and 0.789 (validation), along with the highest accuracy, sensitivity, and F1 score. SHAP analysis identified TG, ALT, GGT, BMI, and UA as the top predictors.	Conducted at a single hospital site. Moderate sample size and lack of external dataset for validation. Some predictors were unavailable due to incomplete clinical records.	Dataset: Survey + Clinical dataset (Chinese T2DM population)

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
24	On the role of explanations in machine learning prediction of nonalcoholic steatohepatitis	To evaluate ML models for NASH detection and study the effect of SHAP-generated and verbal explanations on physician trust and usability.	Decision Tree, Random Forest, XGBoost, CatBoost, SVM, Naive Bayes, KNN, MLP & SHAP + tailored verbal descriptions	The Random Forest model showed strong predictive performance (F-score ~0.69, Accuracy ~0.75, AUC ~0.72), outperforming Decision Trees. In a study with 11 clinicians, explanations modestly increased trust and usability. Some explanations matched medical knowledge, while others revealed model flaws.	Dataset small and filtered by strict inclusion criteria. Clinician survey (n = 11) limits generalizability. Explanations only tested with SHAP, no comparative testing with other XAI tools.	Publicly available UCI machine learning datasets, commonly used for benchmarking predictive models.
25	A machine learning based framework to identify and classify non-alcoholic fatty liver disease in a large-scale population	The study aimed to develop and validate machine learning models for accurate screening of non-alcoholic fatty liver disease (NAFLD) in a large population.	The research utilized LASSO for feature selection and four machine learning algorithms, with XGBoost emerging as the best-performing classifier.	XGBoost demonstrated the best performance with an accuracy of 0.880, precision of 0.801, recall of 0.894, F-1 score of 0.882, and AUC of 0.951, while also ranking the importance of covariates for NAFLD.	While not explicitly stated in the provided abstract, a potential limitation could be the reliance on physical examination data and questionnaires, which may not capture all relevant factors for NAFLD diagnosis.	Large-scale national physical examination dataset of 304,145 adults with questionnaire and physical measurement data related to NAFLD risk factors.

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
26	Explainable machine learning model for predicting the risk of significant liver fibrosis in patients with diabetic retinopathy	To develop an explainable machine learning model capable of predicting the risk of significant liver fibrosis in patients with diabetic retinopathy, thereby enabling early intervention and personalized treatment planning.	Machine learning model: Logistic Regression (LR) with SHAP for interpretability; likely implemented using standard ML libraries such as scikit-learn or R's caret.	The study included 5,364 participants, with 39.45% showing significant liver fibrosis. The Logistic Regression model performed best, achieving an AUC of 0.867 and an F1 score of 0.749.	No limitations were mentioned.	National Health and Nutrition Examination Survey (NHANES) 2005-2008 cohort
27	Explainable AI for Enhanced Interpretation of Liver Cirrhosis Biomarkers	To enhance the interpretability of liver cirrhosis biomarker analysis using explainable artificial intelligence (XAI), aiming for non-invasive diagnosis and improved clinical understanding.	The tech stack includes Explainable AI algorithms, extreme gradient boosting, and Shapley additive explanations for feature importance and visualization.	The paper presents a novel approach to liver cirrhosis diagnosis, leveraging Explainable AI to overcome limitations of previous works and improve diagnostic accuracy.	While the study addresses neglected aspects of liver cirrhosis assessment, it may be limited by its focus on serum biomarkers and the specific dataset of primary biliary cirrhosis cases.	Primary Biliary Cirrhosis (PBC)

No	Title	Objectives	Tech Stack	Results	Limitations	Dataset
28	Life's Crucial 9 and NAFLD from association to SHAP-interpreted machine learning predictions	To evaluate the association between Life's Crucial 9 (LC9)-a cardiovascular health score-and NAFLD risk, and to build interpretable machine learning models for NAFLD prediction.	Statistical tools: Weighted Logistic Regression, Cox Proportional Hazards, Kaplan-Meier curves, Restricted Cubic Spline analysis; ML model: Support Vector Machine (SVM) with SHAP for explainability.	Found an inverse relationship between LC9 scores and NAFLD risk/mortality. The SVM model performed well, with SHAP identifying LC9 components as top predictors. Lifestyle metrics (exercise, sleep, BMI) were clinically valuable in NAFLD risk prediction.	The paper does not explicitly mention limitations. However, potential limitations could include the retrospective nature of the study, reliance on NHANES data, and possible confounding factors not accounted for in the analysis.	National Health and Nutrition Examination Survey (NHANES) dataset from 2007 to 2018.
29	Non-invasive biomarkers for liver inflammation in non-alcoholic fatty liver disease: present and future	The paper aims to review and discuss non-invasive biomarkers for assessing liver inflammation in non-alcoholic fatty liver disease (NAFLD), highlighting both current methods and future possibilities.	No direct model implemented; study reviews imaging techniques (ultrasound, CT, MRI), laboratory biomarkers, and AI/ML-based diagnostic platforms from prior works.	The study highlights inflammation as a key driver of fibrosis in NAFLD but is difficult to assess due to its fluctuating nature and weak correlation with standard liver markers. While histology remains the gold standard, it has notable practical limitations.	The paper highlights the challenges in assessing liver inflammation in NAFLD, including the dynamic nature of inflammation and its poor correlation with liver biochemical markers. Additionally, the standard tool of liver histology is noted to have substantial sampling, intraobserver, and interobserver variability.	1.Optum Administrative Claims Dataset 2.Biopsy Cohorts 3.Public Reference Cohorts (Mentioned in Model Validation) 4.Commercial Biomarker Panels.

2.7 Matrix Decomposition-Based Literature Analysis

Table 2.2: Matrix Decomposition-Based Literature Analysis

No	Title	Objective	Tech Stack	Dataset	Results	Limitations	Features
1	A Deep Matrix Completion Method for Imputing Missing Histological Data	To impute missing histological data using deep matrix completion methods.	Deep Matrix Completion (DMC), Neural Networks, Matrix Factorization	Private dataset (210 breast cancer patients)	Achieved AUC = 0.793; improved imputation accuracy	Small dataset size; limited generalization	Deep learning-based matrix completion; improved feature recovery
2	Missing Value Imputation With Low-Rank Matrix Completion in Single-Cell RNA-Seq Data	To develop a low-rank matrix completion method for imputing missing values in scRNA-seq data.	Low-Rank Matrix Completion, Gauss-Newton Optimization	Human ESC (GSE75748), Mouse brain datasets	Outperformed scImpute, MAGIC, DeepImpute	Poor performance when missing rate $\geq 70\%$	Handles cell heterogeneity; robust low-rank structure
3	Meta-learning-based Inductive Logistic Matrix Completion	To solve data scarcity and few-shot learning problems using matrix completion.	Inductive Logistic Matrix Completion (ILMC), Meta-learning	KinaseDB (drug-target interaction dataset)	AUC = 0.94 in cross-validation	Assumes structured interaction data; domain dependency	Meta-learning integration; handles few-shot scenarios
4	Non-Linear Matrix Completion	To extend matrix completion to nonlinear relationships in data.	IMC, NIMC, Deep IMC (DIMC), Neural Networks	Synthetic and real-world datasets	Better performance on nonlinear data using NIMC/DIMC	High computational cost; complexity in tuning	Nonlinear activation functions; deep learning extension

Table 2.2 provides a comparison of matrix decomposition-based approach for missing data imputation and matrix completion in different application areas. The studies show that traditional low rank matrix completion methods have been extended up to advanced deep learning, meta-learning and non-linear matrix completion methods. In the first study, the Deep Matrix Completion (DMC) approach is combined with neural networks and matrix factorization to impute missing historical biomedical data, obtaining better imputation accuracy with an AUC of 0.793. The second study focuses on the study of single-cell RNA sequencing data, and finds that low-rank matrix completion with Gauss-Newton optimization is more effective than the current imputation methods like scImpute, MAGIC and

DeepImpute. Study 3 proposes a novel meta-learning-based Inductive Logistic Matrix Completion (ILMC) framework that is used to tackle the few-shot learning and data scarcity issues and obtains a high AUC of 0.94 for drug-target interaction data under 5-fold cross-validation. Lastly, the 4th study introduces the concepts of NIMC and DIMC models to accommodate nonlinear data structures, and shows their benefits when applied to complex datasets, even though they involve more complexity of computation and parameter tuning. In conclusion, the studies suggest that the use of deep learning, meta-learning, and nonlinear modeling can substantially improve the performance of matrix completion and its adaptability, albeit with certain limitations and considerations, such as scalability, computational complexity, and generalizability.

2.8 Dataset Analysis Table

Table 2.3: Overview of NAFLD Datasets

Dataset	Short Description	Access
Non-alcoholic fatty liver dataset	Clinical panel (113×62) with BMI, waist, lipids, enzymes, CAP, fibrosis	Public (Kaggle dataset, open access)
Non-alcohol fatty liver disease (NAFLD) part 1	Large cohort (17k×10) with demographics, BMI, follow-up, event status	Public (Kaggle dataset, open access)
Indian Liver Patient Dataset	This dataset contains records of 416 patients diagnosed with liver disease and 167 patients without liver disease. There are 10 variables per patient: age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos in the 583 patient records.	Public (UC Irvine Machine Learning Repository, open access)
Non-alcoholic fatty liver disease	Rich clinical dataset (605×62) with labs, comorbidities, histology, outcomes	Public (Kaggle dataset, open access)
NAFLD collected Dataset	Small clinical dataset (114×13) with BMI, FBS, lipids, enzymes, NAFLD risk	Collected by capstone team

Chapter 3

Proposed System for Non-Alcoholic Fatty Liver Disease Risk Prediction

The chapter outlines the general approach taken in the design, implementation and assessment of an interpretable machine learning architecture to predict the risk of Non-Alcoholic Fatty Liver Disease (NAFLD). The methodology will be based on a structured workflow that encompasses data acquisition and preprocessing, then model training, evaluation and explainability analysis. A number of training and testing processes were developed to support differences in the availability of data, and dimensionality of features. These workflows allow model development with both real collected clinical data and large-scale public datasets as well as cross-dataset generalization analysis. It is aimed that the proposed framework may provide a precise, interpretable and clinically significant NAFLD risk prediction system.

3.1 System Overview

The proposed system is based on an end-to-end machine learning pipeline that is aimed at predicting NAFLD risks with high accuracy, interpretability, and clinical meaning. The framework combines heterogeneous datasets of the population together with a local dataset of real-world clinical data in a single unified feature space. Because of large missing values due to feature heterogeneity, matrix decomposition-based completion is used before class balancing. Multiple machine learning models and an ensemble classifier are then trained on the final processed dataset. SHAP is used to provide model explainability to make the model transparent and clinically

usable. In contrast to the earlier multi-dataset detailed experimental design, all the evaluations are performed in this study based on one final dataset to represent the real deployment environment.

3.2 Methodology

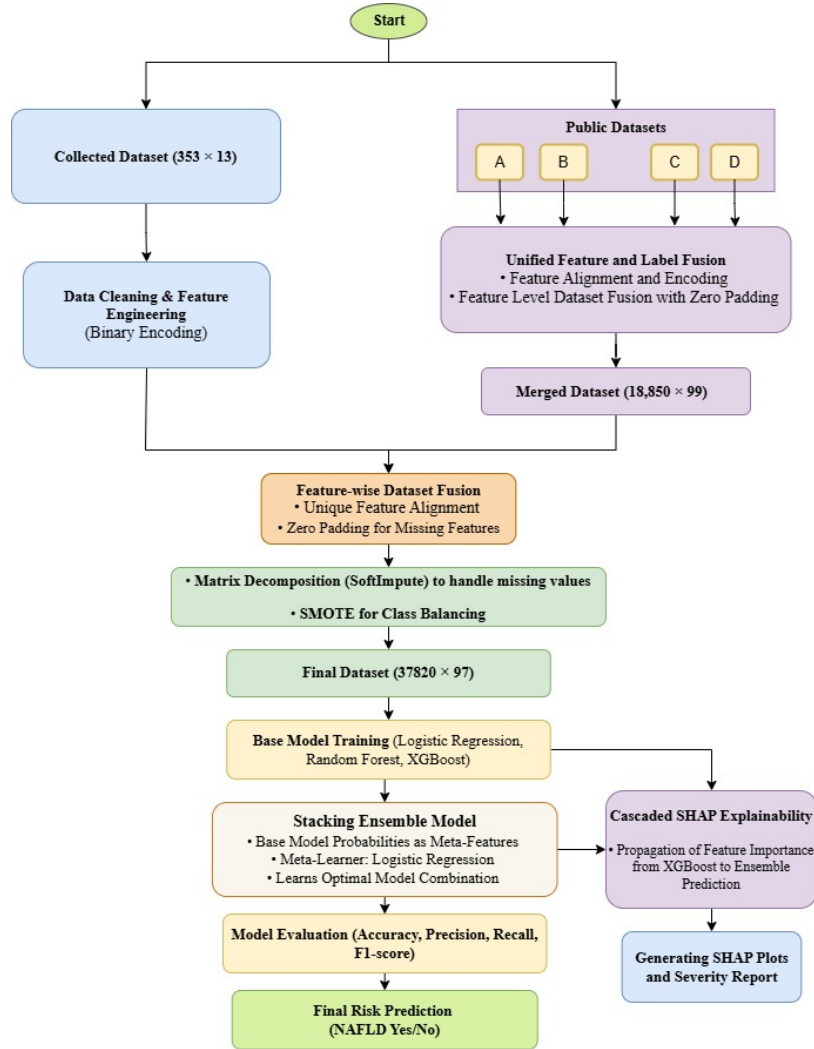


Figure 3.1: Methodology

Figure 3.1 shows the entire end-to-end workflow of the proposed explainable fusion based NAFLD risk prediction system. The pipeline takes two data sources, initially a local collected clinical dataset of 353 records of patients containing 13

features, and secondly four publicly available NAFLD datasets labeled as A-D, namely the Non-Alcoholic Fatty Liver Disease dataset, the NAFLD and NAFLD1 dataset, and Indian Liver Patient Dataset (ILPD).

The public datasets are initially processed with unified feature and label fusion, which consists of feature alignment, feature-encoding, and dataset integration at the feature-level with zero-padding of missing variables. This generates a joint community dataset of 18,850 samples having 99 aligned features. Concurrently the locally gathered dataset is processed by data cleaning and binary encoding which will maintain the real-world clinical traits of the data and remain compatible with the unified space of features. The two streams are then combined by feature-wise dataset fusion where unique features are matched and absent features are structurally padded to create one high dimensional matrix. To deal with large-scale missingness due to heterogeneous feature availability, the low-rank matrix decomposition with SoftImpute is used, and the class balancing based on SMOTE is used to reduce the class imbalance. This leaves the final modeling dataset of 19,203 samples and 97 features.

It is also important to note that before applying SMOTE, the dataset was highly imbalanced (only 293 samples in Class 0). After applying SMOTE, the dataset became balanced (18,910 samples per class, 37,820 samples total), which contributed to improved model learning and more reliable classification results.

Then the Logistic Regression, Random Forest and XGBoost classifiers are trained on the processed dataset. These base models are then used as inputs to a stacking ensemble framework, in which a Logistic Regression meta-learner is used to combine the results of the base models to make the final prediction. The mentioned stacking method enables the system to learn the best combination of model predictions and enhances better robustness and generalization. Accuracy, precision, recall, and F1-score are used to evaluate model performance and the system then generates the final NAFLD risk prediction in the binary form.

At the same time, model explainability is incorporated using a cascaded SHAP approach. SHAP values are first computed for the trained XGBoost model to estimate feature-level contributions. The stacking ensemble is then applied to SHAP analysis to establish the impact of the base model outputs on the end result. All these explanations are combined to spread the feature significance of XGBoost to the ensemble prediction to facilitate feature-wise interpretation of the final model.

According to this analysis, SHAP feature importance plots, beeswarm plots, dependence plots, and patient-level waterfall plots are created as well as a SHAP-based severity report. These explanations make the model understandable about its decision making, and can be used to make NAFLD risk assessment in a clinically comprehensible way.

3.3 Requirement Analysis

This subsection provides the functional, software, and hardware specifications of the proposed explainable fusion-based NAFLD risk prediction system, which guarantees reproducibility, scalability, as well as compatibility with cloud-based clinical analytics pipelines.

3.3.1 Functional Requirements

To overcome the structural differences between datasets, the proposed system combines various publicly accessible NAFLD datasets with locally obtained clinical data into a single feature space by aligning and encoding the features and zero-padding them. Missing numerical values are replaced with low-rank matrix decomposition with the SoftImpute algorithm and imbalance of classes is addressed by SMOTE oversampling to generate a balanced dataset. The system takes input data, then through machine learning models, risk prediction of NAFLD is made, and predictive performance is estimated using common classification measures such as accuracy, precision, recall, F1-score. The system also generates explainability output through a cascaded SHAP framework to achieve transparency and clinical interpretability and allow the analysis of the global feature importance, a patient-level explanation of NAFLD risk, and severity reports that quantify the contribution of any particular clinical feature to the predicted NAFLD risk.

3.3.2 Software and Platform Requirements

FancyImpute library with SoftImpute algorithm is used to complete missing values. Also, the SMOTE technique is used to correct the class imbalance. The SHAP framework provides the means of feature attribution analysis and explainable predictive information, which is attained by model interpretability.

3.3.3 Hardware and Runtime Configuration

Google Colab CPU runtimes were mainly used to perform the model development and evaluation. In computationally intensive tasks, namely, stacking ensemble inference and SHAP-based explanation generation, a T4 GPU runtime was used to perform the computations faster.

Chapter 4

Design and Development

4.1 Dataset

Two main sources of primary data are used in this study, including various publicly accessible NAFLD datasets and a local real-world clinical dataset. The publicly available datasets are different in terms of the size of the cohort, population structures, and the composition of features, which all present a wide range of demographic, anthropometric, biochemical, and metabolic measures. The dataset that is collected is actual records of patients collected under the supervision of physicians and it represents the Bangladeshi population and includes regularly accessible clinical and laboratory measurements.

4.1.1 Collected Clinical Dataset

The collected clinical dataset includes 353 actual patients records, which were gathered directly with the help of physicians based on hospital data. The dataset contains 13 clinically significant characteristics, e.g., age, body mass index (BMI), systolic blood pressure, diastolic blood pressure, liver enzymes (ALT, AST), lipids profile components (triglycerides, HDL, LDL), fasting blood glucose, and diabetes condition. These variables were chosen according to the available medical literature and the recommendations given by experts since they are known risk factors of NAFLD. There was little preprocessing that was used so as to maintain the real-world properties and clinical authenticity of the data. A screenshot of the collected dataset can be seen in Figure 4.1.

	Age	Height	Weight	BMI	FBS	ALT	AST	LDL	HDL	Triglycerides	Cholesterol	Diabetes Status (Yes/No)	label
2	25	5'5	67	24.57993082	110	64	45	140	38	190	210	no	yes
3	27	5'6	78	27.75487645	112	65	48	132	35	220	230	no	yes
4	32	5'4	85	32.16559167	125	72	58	145	33	250	240	no	yes
5	29	5'8	90	30.16874546	135	80	62	150	32	180	200	yes	yes
6	31	5'2	70	28.2258629	108	60	46	128	40	190	250	no	yes
7	45	5'7	95	32.80247149	140	85	65	160	30	270	198	yes	yes
8	24	5'5	75	27.51484793	110	58	42	120	42	175	220	no	yes
9	25	5'3	80	31.24218896	120	68	50	135	36	210	230	no	yes
10	24	6'1	114	33.15825735	100	78	45	107	44	108	216	no	yes
11	35	5'5	78	28.61544185	118	58	46	130	38	180	210	no	yes
12	42	5'7	84	29.00429058	132	64	52	140	35	195	225	yes	yes
13	39	5'6	82	29.17820344	125	60	50	135	36	190	218	yes	yes
14	26	5'4	60	22.70512354	85	22	21	95	55	110	175	no	no
15	45	5'8	90	30.16874546	140	72	58	150	32	205	235	yes	yes

Figure 4.1: Screenshot of the Collected Dataset

4.1.2 Public NAFLD Datasets

Four datasets on NAFLD available publicly were included in Kaggle and the UCI Machine Learning Repository to make a bigger and more diverse dataset. These datasets are quite different in terms of features composition, size of cohort, and population characteristics.

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	ALP	ALT	AST	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Patient's Liver Disease(Yes/No)
2	65	0	0.7	0.1	187	16	18	6.8	3.3	0.9	1
3	62	1	10.9	5.5	699	64	100	7.5	3.2	0.74	1
4	62	1	7.3	4.1	490	60	68	7	3.3	0.89	1
5	58	1	1	0.4	182	14	20	6.8	3.4	1	1
6	72	1	3.9	2	195	27	59	7.3	2.4	0.4	1
7	46	1	1.8	0.7	208	19	14	7.6	4.4	1.3	1
8	26	0	0.9	0.2	154	16	12	7	3.5	1	1
9	29	0	0.9	0.3	202	14	11	6.7	3.6	1.1	1
10	17	1	0.9	0.3	202	22	19	7.4	4.1	1.2	2
11	55	1	0.7	0.2	290	53	58	6.8	3.4	1	1
12	57	1	0.6	0.1	210	51	59	5.9	2.7	0.8	1
13	72	1	2.7	1.3	260	31	56	7.4	3	0.6	1
14	64	1	0.9	0.3	310	61	58	7	3.4	0.9	2
15	74	0	1.1	0.4	214	22	30	8.1	4.1	1	1

Figure 4.2: Screenshot of the ILPD Dataset

The Indian Liver Patient Dataset (ILPD) consists of 583 records and 11 features, most of which are demographic variables and liver-related function variables including bilirubin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase, albumin, and total proteins. There is a widespread application of this dataset in liver disease classification tasks, and it offers a small number of core biochemical indicators[35]. A snippet of the ILPD dataset can be seen in Figure 4.2.

NAFLD data has 605 samples, 61 features which provide a more detailed clinical profile. It contains specific biochemical, metabolic, and anthropometric variables, including lipid profile values, glucose and insulin, fibrosis-related variables, and inflammatory variables, which allows more detailed representation of the features in NAFLD risk modeling [36].

The NAFLD1 dataset is the biggest dataset employed in the research that contains 17,549 samples and 7 features. It is concerned with the most important clinical indicators, such as age, BMI, liver enzyme measurements, and diabetes status. It has limited dimensional feature representation, but a high sample size gives it an advantage in population-level learning [37].

The Non-Alcoholic Fatty Liver dataset contains 113 samples with 61 features, overlapping substantially with the NAFLD dataset in terms of clinical variables but representing a different patient cohort. This dataset contributes additional diversity in population characteristics and feature distributions [38].

Age	Weight	Height	BMI	Waist	ALT	AST	HDL	Uric Acid	Triglycerides	Platelets	LDL	Cholesterol	RBC	WBC	FBS
34	96	186	27.7	102	28	21	47	6.2	56	347	98	153	5.46	9220	102
48	88	173	29	103	17	24	58	7	78	239	106	170	7	6892	104
52	88	173	29	103	26	24	55	6.6	146	339	223	287	5.35	4820	99
68	88	173	29	103	19	25	42	6.6	74	239	104	162	7	6892	98
68	88	173	29	103	14	24	53	5	73	239	145	227	7	6892	100
46	88	173	29	103	17	22	51	5.8	91	269	94	155	5.04	4200	85
39	75.4	184	22.3	88	16	17	40	5.3	99	130	88	155	5	4600	99
34	88	173	29	103	118	74	42	6	197	239	118	183	7	6892	98
40	58	160	22.7	103	48	34	33	5.8	112	239	151	252	7	6892	99
69	94	175	30.7	105	19	23	55	5.5	113	188	90	168	4.9	5200	102
38	91	160	35.5	118	34	29	55	4.5	71	246	97	170	4.6	5500	103
36	53	155	22.1	74	12	15	40	3.6	84	239	90	139	7	6892	95
39	88	173	29	103	11	19	55	5.6	73	239	139	208	7	6892	96
39	88	173	29	103	118	74	42	6	197	239	118	183	7	6892	98
51	88	173	29	103	21	28	34	6	50	239	108	172	7	6892	107
37	88	173	29	103	22	27	45	3.1	65	239	107	171	7	6892	85
58	68	163	25.6	85	39	30	60	5.1	71	190	113	214	4.6	5600	106
47	88	173	29	103	11	18	45	4.8	55	239	104	176	7	6892	97
41	88	173	29	103	26	23	40	4	200	274	115	195	5.71	6630	88
44	88	173	29	103	118	74	42	6	197	239	118	183	7	6892	98
55	66	169	23.1	82.4	14	16	44	4.5	144	263	137	210	5.13	9060	171
44	88	173	29	103	118	74	42	6	197	239	118	183	7	6892	98
50	101.4	174	33.5	111.6	118	166	39	8.2	240	223	113	200	5.79	9090	97
46	107.6	182	32.5	110.9	41	34	40	4.4	193	184	121	200	6.49	9360	117

Figure 4.3: Screenshot of the Combined Dataset

Since none of the individual public datasets alone provided both sufficient sample size and comprehensive feature coverage, a feature-level dataset fusion strategy was employed. All unique features across the four datasets were aligned and standardized, with missing features handled through zero-padding to ensure uniform dimensionality. Label formats were also harmonized into a single binary NAFLD outcome. The resulting merged public dataset consists of 18,850 samples with 99 aligned features, covering a wide range of demographic, anthropometric,

biochemical, metabolic, and fibrosis-related variables. Next, the 353 data rows were added to this dataset. Figure 4.3 shows the final combined dataset.

Together, these datasets support comprehensive experimentation across real-world, high-dimensional, and clinically constrained scenarios.

4.2 Dataset Preprocessing and Feature Alignment

Since there was a heterogeneity in the datasets, a dataset fusion strategy with feature utilization was applied. All features over sets were standardized to a unique feature space and missing feature values were generated at positions of data unavailability. Forms of labelling were consolidated to binary NAFLD results. The clinical data set collected was preprocessed, encompassing the unit normalization and nominal encoding, to not lose the actual value of a clinical feature but to allow the integration of the united feature set.

4.3 Matrix Decomposition for Missing Value Completion

The combined dataset had a large number of missing values which were caused by feature heterogeneity. In order to solve this problem without losing the relationship between features, the SoftImpute algorithm was used to perform low-rank matrix decomposition. This approach is used to estimate missing values using latent feature correlations and retains underlying data structure better than the traditional imputation methods. Before balancing of the classes, matrix decomposition was done on numerical features only. SoftImpute operates on filling-in the values that are missing, seeking to fit the known values in the most likely manner possible and ensuring that the matrix is simple and low-rank. The SoftImpute method's formula is given in (1), adapted from [34].

$$\min_M \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|M\|_* \quad (4.1)$$

Here, X is the incomplete data matrix, $P_\Omega(\cdot)$ is the projection operator onto observed values, $\|\cdot\|_F$ denotes the Frobenius norm, $\|M\|_*$ denotes the nuclear norm, and λ controls the degree of rank shrinkage. This formulation allows estimating the

missing data based on shared low-dimensional structure, making it more suitable for heterogeneous clinical data than traditional imputation methods.

4.4 Class Imbalance Handling and Dataset Splitting

After completing the matrix, there was class imbalance in the dataset. Synthetic Minority Oversampling Technique (SMOTE) was used on the achieved feature matrix to evenly sample NAFLD and non-NAFLD classes. This balanced dataset was further split into training, validation and test sets with stratified 70:15:15 to avoid data leakage and provide reproducibility.

4.5 Model Training and Ensemble Learning

Three machine learning models including Logistic Regression, random forest, and XGBoost were trained on the final balanced dataset. Where necessary, feature scaling was used to make model training stable. The predicted probabilities of the base models were used as meta-features in stacking ensemble structure after the base model was trained. A Logistic Regression model was employed as the meta-learner to combine the outputs of the base models and produce the final prediction. All models were evaluated on the same training, validation and test splits with the same performance metrics to compare the predictive performance fairly and consistently across all models.

4.6 Explainability Using SHAP

SHAP (SHapley Additive exPlanations) was used to increase model transparency and clinical interpretability in a cascaded explainability implementation. First, SHAP values were calculated for the trained XGBoost model to estimate feature-level contributions to its predictions. SHAP analysis was then used on the stacking meta-model to quantify the impact of the base model outputs on the final ensemble prediction. Next, the two explanations were merged to pass on the contributions of features that XGBoost makes to the ensemble model and, therefore, it is possible to estimate the effects of each original clinical feature on the ultimate prediction of NAFLD risk. The SHAP feature importance and beeswarm plots were used

to examine global model behavior, and show the general impact and direction of clinical features on risk of disease. Non-linear relationships between significant predictors and model outputs were also investigated by creating feature dependence plots. To interpret predictions at a patient level, SHAP waterfall plots were applied to reveal how specific feature values contribute to risk compared to a baseline prediction. Besides that, a SHAP-based severity report was generated to summarize the contribution of features, probability of risk being predicted, and confidence of each case. This global and local set of explanations has offered clinical value of how the ensemble model makes decisions.

4.7 Web Application Overview

The proposed system is deployed as a full-stack web application named *LiverGuard*, designed to provide real-time NAFLD risk prediction using machine learning. The application integrates a React-based frontend, a Node.js backend with MySQL database support, and a Python Flask API for model inference.

The system allows users to input routinely available clinical parameters such as age, BMI, liver enzyme levels, lipid profile, and diabetes status. These inputs are processed and forwarded to the machine learning API, where trained models generate predictions along with probability scores and contributing factors.

To ensure practical usability, a reduced feature set of 12 clinically accessible attributes is used for deployment. Machine learning models, including Logistic Regression, Random Forest, XGBoost, and a stacking ensemble, were retrained using this reduced dataset and deployed within the system.

The application provides:

- NAFLD risk classification
- Probability score
- Key contributing factors
- Clinical recommendation

This deployment enables a non-invasive, fast, and user-friendly decision-support tool for early NAFLD screening.

Chapter 5

Result Analysis

5.1 Model Performance Evaluation

In this section, the proposed explainable fusion framework of NAFLD risk prediction is thoroughly evaluated. Class-wise Precision, Recall, F1-score, overall Accuracy, Receiver Operating Characteristic (ROC) curves, and SHAP-based interpretability are used to measure the performance of Logistic regression, random forest, XGBoost, and the soft-voting Ensemble model.

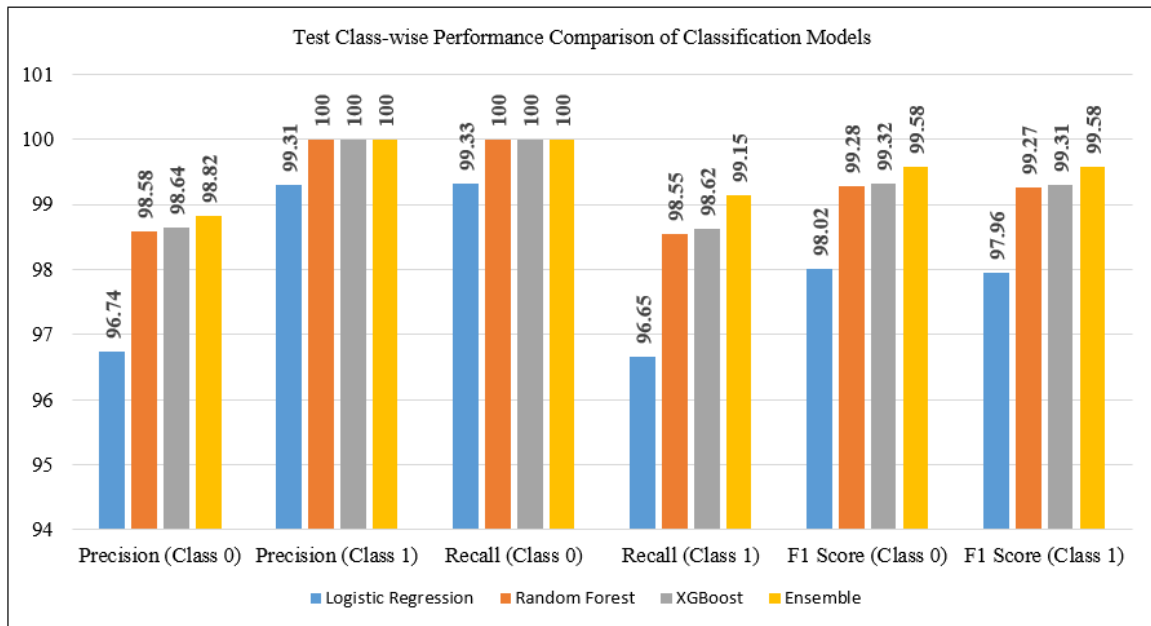


Figure 5.1: Precision, Recall & F1-score Comparison of Models on the Test Dataset

Figure 5.1 shows the Precision, Recall, and F1-score of the Logistic Regression,

Random Forest, XGBoost and Ensemble model on the test data by each class.

From the figure, it is clear that XGBoost and the Ensemble model perform the best overall, achieving near-perfect scores across almost all metrics. In particular, for Class 1 (NAFLD), Random Forest, XGBoost, and the Ensemble achieve 100% precision and recall, indicating that these models are highly effective in correctly identifying positive NAFLD cases without misclassification.

On the contrary, Logistic Regression exhibits the worst results in all aspects, particularly in Recall (Class 1), where it reduces to approximately 96.65. This indicates that Logistic Regression has a higher chance of missing positive NAFLD cases (false negatives), which is a serious limitation in healthcare applications. This reduced performance of the Logistic Regression can be explained by its linear nature, which is less able to represent the nonlinear relationships that exist in clinical data.

For Class 0 (non-NAFLD), all models perform relatively well, but again Ensemble and XGBoost slightly outperform others, achieving the highest F1-scores. The Ensemble model has constant and good performance in both classes, which means that it can integrate strengths of various models and eliminate weaknesses of each model.

Overall, Figure 5.1 demonstrates that ensemble learning and boosting techniques significantly improve classification performance, particularly for detecting NAFLD cases, making them more suitable for real-world clinical applications.

The comparison of Precision, Recall and F1-score of Logistic Regression, Random Forest, XGBoost and the Ensemble model on the validation dataset in terms of their class-wise performance is provided in Figure 5.2. The results follow a similar trend to the test dataset, confirming the consistency and generalization ability of the models.

From the figure, XGBoost and the Ensemble model again demonstrate the best performance, achieving nearly perfect scores across all evaluation metrics for both Class 0 (non-NAFLD) and Class 1 (NAFLD). The Ensemble model has the best overall F1-scores, with a maximum of approximately 99.58% Class 1, which means that there is a good balance between the precision and the recall. These high values show that the Ensemble model is highly reliable in identifying both positive and negative cases with minimal misclassification. Random Forest also performs very strongly, with performance values very close to XGBoost and Ensemble. This

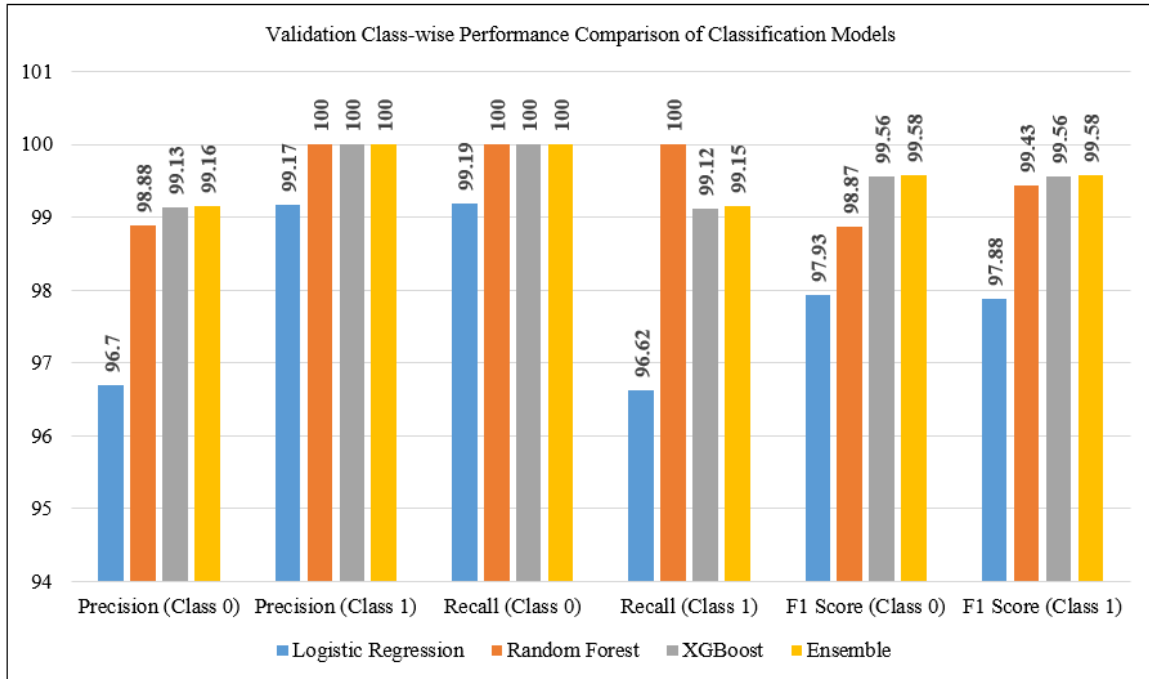


Figure 5.2: Precision, Recall & F1-score Comparison of Models on the Validation Dataset

means that tree-based models are very good in this dataset since they can establish nonlinear associations and feature interactions among clinical biomarkers.

On the other hand, Logistic Regression again shows the lowest performance, particularly in Recall (Class 1), which is around 96.62%. This lower recall indicates that Logistic Regression has a slightly higher probability of missing NAFLD-positive cases. This comparatively poor performance is due to the fact that Logistic Regression is based on the assumption of a linear relationship between input variables and the variable to predict, which might not be entirely true to the complex interdependence of metabolic and clinical variables that contribute to the prediction of NAFLD.

The similarity of performance trends between the test dataset (Figure 5.1) and the validation dataset (Figure 5.2) suggests that the models are not overfitting and have strong generalization capability. The stable performance of XGBoost and the Ensemble model across both datasets indicates that these models are robust and reliable for real-world prediction tasks.

Overall, Figure 5.2 confirms that ensemble learning and boosting techniques

consistently outperform single linear models, providing more accurate and stable predictions for NAFLD risk detection.

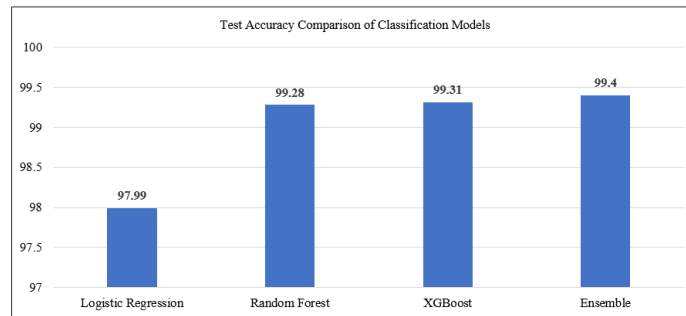


Figure 5.3: Accuracy Comparison across Models on the Test Dataset

From Figure 5.3 (Test Dataset), it can be observed that the Ensemble model achieves the highest accuracy of 99.40%, followed closely by XGBoost (99.31%) and Random Forest (99.28%). In contrast, Logistic Regression shows the lowest accuracy at 97.99%. This indicates that while Logistic Regression performs reasonably well, it is less effective compared to the more advanced models.

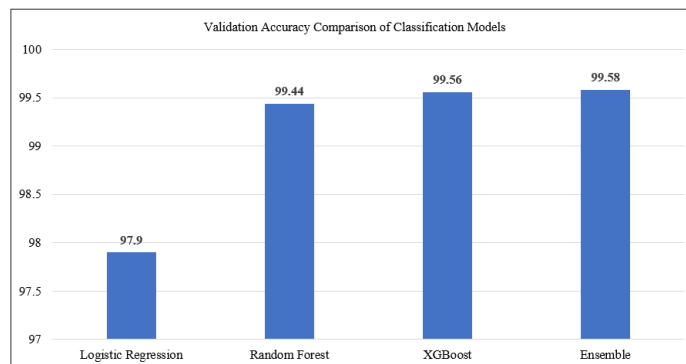


Figure 5.4: Accuracy Comparison across Models on the Validation Dataset

A similar pattern is observed in Figure 5.4 (Validation Dataset), where the Ensemble model again achieves the best performance with 99.58% accuracy, followed by XGBoost (99.56%) and Random Forest (99.44%), while Logistic Regression remains the lowest at 97.9%. The consistency of these results across both datasets indicates that the models have strong generalization ability and are not overfitting.

The superior performance of XGBoost and Ensemble models can be attributed to their ability to capture complex nonlinear relationships and interactions between

clinical features, which are common in medical datasets such as NAFLD. XGBoost is based on a boosting mechanism, which continually improves the model performance by adjusting the earlier errors, and the Ensemble model further improves prediction by summing the results of many base models.

On the other hand, the relatively lower performance of Logistic Regression is due to its assumption of a linear relationship between input features and the target variable. This restricts its capability to capture the complicated interactions between biochemical and clinical variables, leading to a relatively lower accuracy.

The small but consistent improvement from Random Forest to XGBoost and finally to the Ensemble model demonstrates the effectiveness of model optimization and fusion techniques. The Ensemble model benefits from combining different learning patterns, leading to improved robustness and predictive performance.

Overall, Figures 5.3 and 5.4 confirm that ensemble learning and boosting-based models provide superior accuracy and reliability, making them more suitable for real-world NAFLD risk prediction tasks compared to traditional linear models.

Table 5.1: Class-Wise Test Performance Metrics Across Classification Models

Models	Accuracy	Precision		Recall		F1 Score	
		Class Label		Class Label		Class Label	
		0	1	0	1	0	1
Logistic Regression	0.9799	0.9674	0.9931	0.9933	0.9665	0.9802	0.9796
Random Forest	0.9928	0.9858	1.0000	1.0000	0.9855	0.9928	0.9927
XGBoost	0.9931	0.9864	1.0000	1.0000	0.9862	0.9932	0.9931
Ensemble	0.9940	0.9882	1.0000	1.0000	0.9880	0.9940	0.9940

Table 5.1 shows the precise numerical values of Accuracy, Precision, Recall and F1-score of each model on the test dataset, which is more detailed as compared to the graphical results.

The Ensemble model has the best overall performance, with an accuracy of 99.40% and the highest F1-score of both classes, which reflects a good balance between the recall and the precision. The XGBoost and the Random Forest follow right behind, with nearly perfect values, particularly Class 1, with precision and recall of 1.0000, showing an excellent detection of NAFLD cases.

Conversely, the lowest performance is recorded by the Logistic Regression especially in recall of Class 1 (0.9665), which means a higher probability of missing

positive cases. This indicates its weaknesses in the modeling of nonlinear relationships.

One of the main findings of the table is that the tree-based and ensemble models are consistently high in performance with both classes, whereas the Logistic Regression is relatively less stable. The slight yet steady enhancement between Random Forest and XGBoost and at last to the Ensemble model serves as an indicator of the advantage of boosting and model fusion methods. Altogether, the table proves that the Ensemble model offers the most optimal and trustworthy performance, whereas Logistic Regression is the least effective among the compared models.

Figure 5.5 shows a comparison of the confusion matrices of Logistic Regression, Random Forest and XGBoost in both validation and test datasets for 5,673 samples each. The matrices are used to demonstrate the accuracy of each model in classifying NAFLD (Class 1) and non-NAFLD (Class 0).

According to the figure, Random Forest and XGBoost have much fewer misclassifications in comparison to Logistic Regression. Logistic Regression reported a comparatively higher number of false negative cases in which NAFLD-positive cases were wrongly predicted as non-NAFLD. This can be seen in the increased off-diagonal values in the Logistic Regression matrices (such as false negatives around 95-96). False negatives are especially important in medical prediction problems since they represent the disease cases missed.

Random Forest and XGBoost, in contrast, demonstrate near-ideal classification rates, and the number of false positives and false negatives in both validation and test sets are zero and very low, respectively. XGBoost is a little better than the Random Forest in that it reduces the false negatives even more, and it is more sensitive to identify NAFLD positives. The diagonal values of all matrices are large, meaning that the majority of predictions are accurate. Nevertheless, the difference between Logistic Regression and Random Forest and XGBoost indicates that tree-based models are more effective in capturing all the complex relationships between the clinical features and thereby make fewer errors during classification.

The other significant observation is that the confusion matrices of validation and test data sets have very similar patterns in each model. It means that the models do not overfit and have consistent performance with various datasets.

In general, the comparison proves that XGBoost and Random Forest can offer more trustworthy results in classification compared to logistic regression, in terms

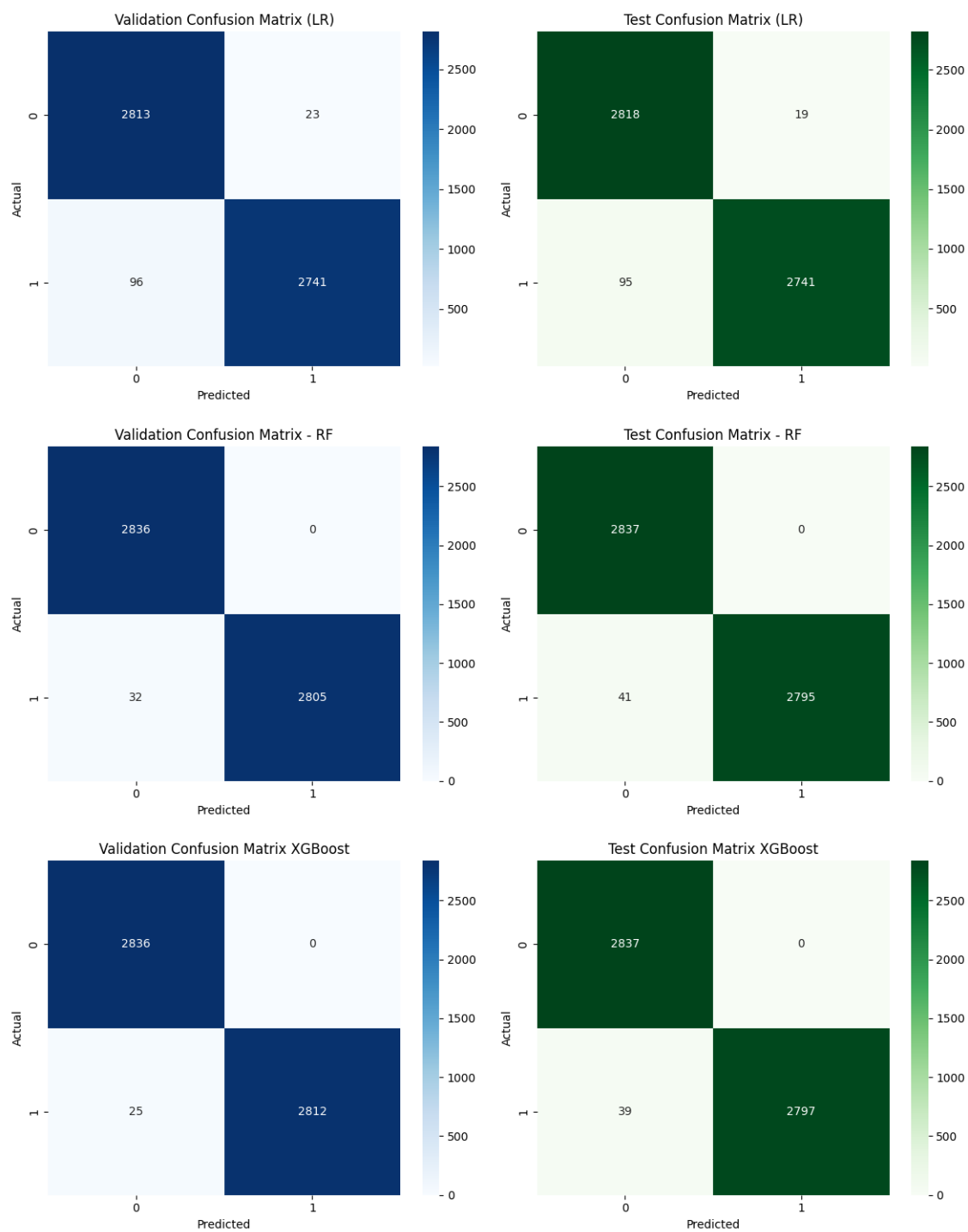


Figure 5.5: Comparison of Confusion Matrices across Models

of reducing false negatives. This renders the tree-based models more appropriate in predicting NAFLD since it is important to accurately detect positive cases and implement early medical intervention.

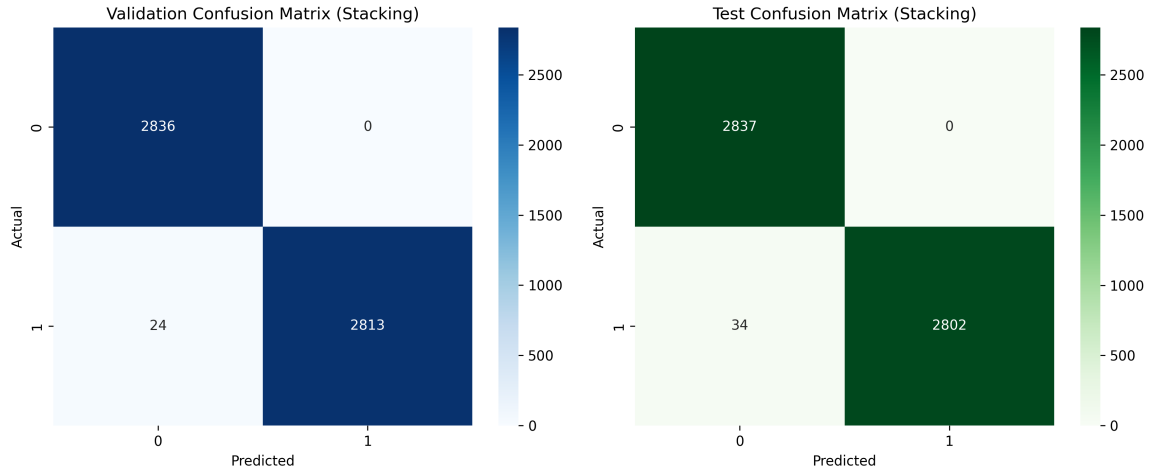


Figure 5.6: Confusion Matrices of the Stacked Ensemble Model

The confusion matrices of stacked Ensemble model on the validation and test datasets of 5,673 samples each are provided in Figure 5.6. The matrices explicitly indicate the better performance of the Ensemble model as compared to the individual models in terms of classification.

Based on the validation confusion matrix, the model is able to correctly classify 2836 non-NAFLD (True Negatives) and 2813 NAFLD (True Positives) and 0 false positives and just 24 false negatives. Likewise, on the test dataset, the Ensemble model accurately predicts 2837 non-NAFLD and 2802 NAFLD, and zero false positives as well as 34 false negatives. The findings imply that the classification accuracy is extremely high and relatively low errors on prediction.

The Ensemble model, as opposed to the single models presented in Figure 5.5, has even fewer false negatives, i.e., less NAFLD-positive patients are wrongly classified as healthy. This increase shows how a combination of several models is effective since the Ensemble approach uses the benefits of Logistic Regression, Random Forest, and XGBoost and mitigates the weaknesses of the respective models.

The next significant finding is a high level of consistency between validation and test confusion matrices. The comparable spread of the prediction outcomes

shows that the Ensemble model is able to exhibit constant performance in various datasets, which proves a high level of generalization.

The fact that the model does not record false positives is also important because it shows that the model will not misclassify healthy individuals as diseased. In clinical practice, it is necessary to reduce the false negative and false positive in order to have reliable decision support.

On the whole, Figure 5.6 proves that the stacking Ensemble model has the best balanced and reliable classification results, making the highest accuracy predictions with the least errors. This illustrates why model fusion methods are beneficial in enhancing predictive performance and robustness of NAFLD risk detection.

5.2 SHAP Explainability

The SHAP beeswarm plot in Figure 5.7 provides a global view of how individual clinical variables influence NAFLD prediction across the dataset. features such as LDL, HDL, Albumin-Globulin Ratio, Gender, Direct Bilirubin, Age, FibroScan score, Triglycerides, and ALT emerge as dominant contributors. Higher LDL levels are associated with large positive SHAP values, indicating a strong push toward NAFLD-positive predictions, whereas HDL generally contributes negatively, suggesting a protective effect. The dispersion of points across SHAP values highlights both population-level importance and patient-specific variability, supporting the model's capacity to capture heterogeneous clinical profiles.

The mean absolute SHAP ranking shown in Figure 5.8 summarizes the overall contribution magnitude of each predictor. LDL is identified as the most influential biomarker, followed by HDL, Albumin-Globulin Ratio, Gender, Direct Bilirubin, Age, and FibroScan score, confirming their dominant role in decision formation. This ranking is consistent with the beeswarm visualization and reinforces the clinical relevance of lipid profiles and liver-stiffness indicators in NAFLD risk stratification.

Figure 5.9 demonstrates a SHAP waterfall plot for a representative patient, where features such as Monocyte count, LDL, and FibroScan score significantly increase the predicted NAFLD probability, while others such as Index and Eosinophil reduce it. This patient-level explanation supports personalized clinical decision-making.

The dependence plot in Figure 5.10 highlights the nonlinear relationship between LDL levels and SHAP values, confirming that elevated LDL contributes sharply to

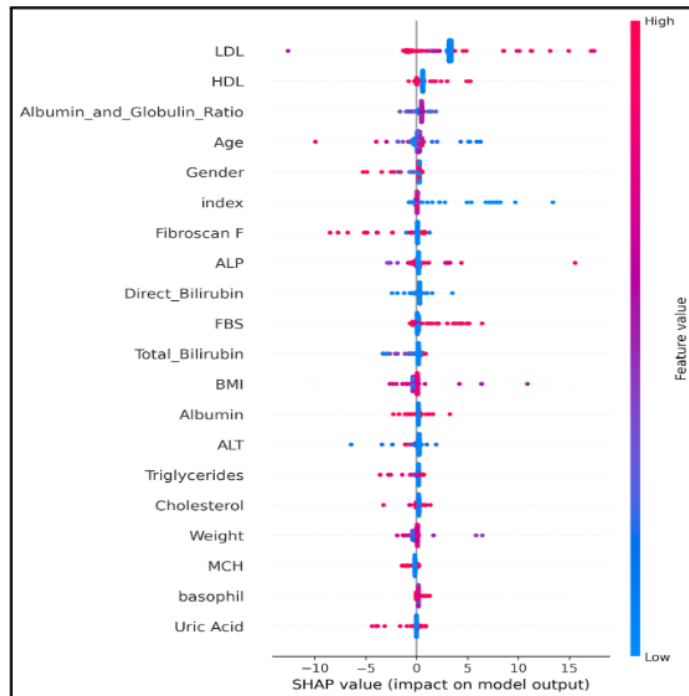


Figure 5.7: SHAP Beeswarm Plot of the Stacked Ensemble Model

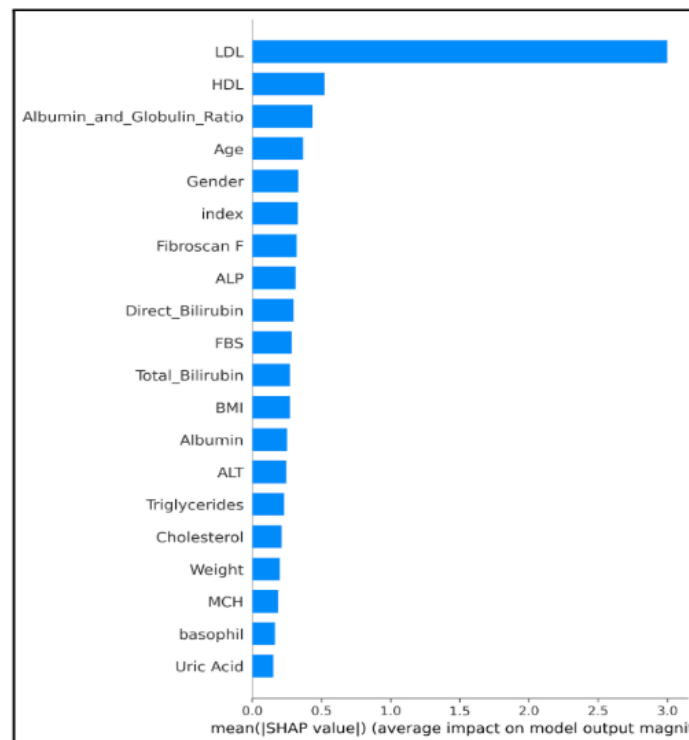


Figure 5.8: SHAP Feature Importance Plot of the Stacked Ensemble Model

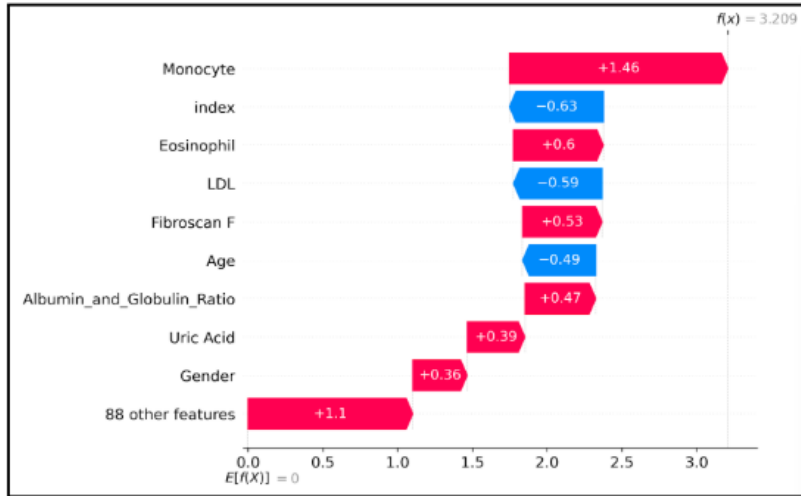


Figure 5.9: SHAP Waterfall Plot of the Stacked Ensemble Model

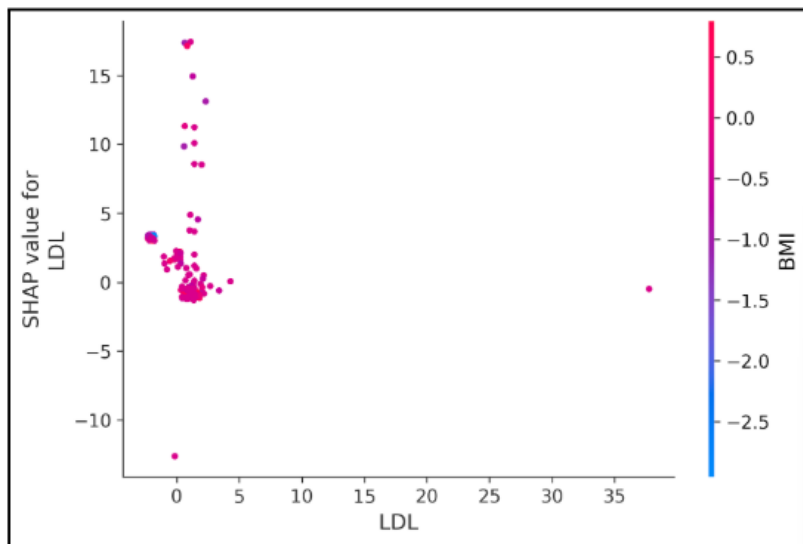


Figure 5.10: SHAP Dependence Plot of the Stacked Ensemble Model

higher NAFLD risk beyond a certain threshold. Such visualization aids clinicians in identifying critical biomarker ranges.

Table 5.2: SHAP Explainability Report of the Stacked Ensemble Model

Predicted Label	Risk Percentage	Top Feature	Top 3 Features	Total SHAP Impact	Severity Level
1	100	Monocyte	[('Monocyte', 1.4615), ('Index', 0.6292), ('Eosinophil', 0.6022)]	9.4097	High
1	99.99	Monocyte	[('Monocyte', 1.4444), ('Cholesterol', 0.7118), ('Fibroscan F', 0.5167)]	8.1651	High
1	87.82	Age	[('Age', 3.9538), ('LDL', 3.7658), ('Uric Acid', 3.1287)]	28.7622	High
1	99.99	LDL	[('LDL', 0.5125), ('Uric Acid', 0.4817), ('Fibroscan F', 0.4348)]	4.3241	High
0	0.33	LDL	[('LDL', 17.439), ('Fibroscan F', 7.7174), ('Age', 6.2863)]	75.5991	Low
1	99.99	LDL	[('LDL', 0.1381), ('Fibroscan F', 0.1035), ('Triglycerides', 0.0695)]	0.9123	High
0	0.92	LDL	[('LDL', 11.2409), ('FBS', 4.4331), ('Uric Acid', 3.7976)]	62.1596	Low
1	100	LDL	[('LDL', 0.9981), ('Fibroscan F', 0.7716), ('Direct_Bilirubin', 0.5903)]	8.0907	High
1	99.98	LDL	[('LDL', 0.8823), ('Fibroscan F', 0.7908), ('Uric Acid', 0.4849)]	6.5528	High
1	99.98	Uric Acid	[('Uric Acid', 0.3969), ('LDL', 0.3515), ('Triglycerides', 0.3449)]	3.4136	High
1	100	LDL	[('LDL', 1.0508), ('Uric Acid', 0.896), ('Fibroscan F', 0.7176)]	7.8503	High
1	100	Uric Acid	[('Uric Acid', 0.9001), ('LDL', 0.8697), ('Fibroscan F', 0.6046)]	7.9768	High
1	99.99	LDL	[('LDL', 1.0642), ('Fibroscan F', 0.6374), ('Gender', 0.4441)]	6.0998	High
1	100	LDL	[('LDL', 0.9013), ('Fibroscan F', 0.5592), ('Albumin_and_Globulin_Ratio', 0.3931)]	5.817	High

Table 5.2 presents a case-wise SHAP explainability report summarizing predicted NAFLD risk, probability score, severity category, and the top contributing features for multiple individuals. The table demonstrates how the model assigns High, Moderate, and Low risk levels based on cumulative SHAP impact and highlights the dominant features driving each prediction. Variables such as Monocyte count, LDL, Age, FibroScan score, Index, and Uric Acid repeatedly appear among the top contributors, indicating their consistent influence across patients. This structured explanation bridges model predictions with clinically interpretable factors, enabling transparent patient-level risk assessment and facilitating physician trust in automated screening systems.

Overall, the experimental results demonstrate that XGBoost and the proposed Ensemble fusion model provide superior predictive performance, achieving near-perfect class discrimination and stable generalization. The ROC analysis confirms

outstanding separability, while SHAP explanations ensure interpretability, an essential requirement for deployment in medical decision-support systems.

Chapter 6

Web Application

6.1 Web Application Implementation

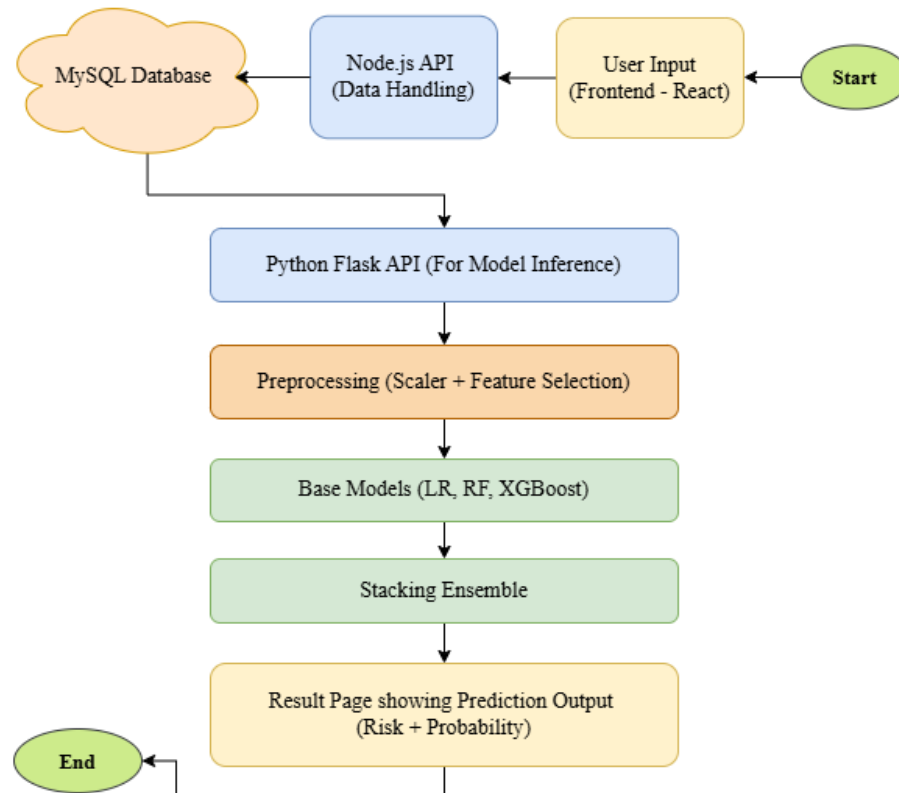


Figure 6.1: Workflow of the Website

The full workflow of the website can be seen in Figure 6.1. The LiverGuard system follows a multi-layer architecture consisting of frontend, backend, and

machine learning components.

- The user inputs patient data through the frontend interface.
- The frontend sends data to the Node.js backend API.
- The backend stores the data in a MySQL database.
- The ML API performs prediction using trained models.
- The prediction results are returned to the frontend and displayed to the user.

6.2 Model Deployment

The deployed models include:

- Logistic Regression
- Random Forest
- XGBoost
- Stacking Ensemble

These models are trained on a reduced 12-feature dataset, ensuring compatibility with real-world clinical input availability. A preprocessing scaler and feature alignment module are applied before prediction.

The stacking ensemble model combines base model predictions to produce the final output.

6.3 User Interface and Functionality

The web application includes several key pages:

- Input Page: Collects patient clinical data with validation and automatic BMI calculation
- Result Page: Displays risk level, probability score, contributing factors, and recommendations
- Performance Page: Shows model evaluation metrics and confusion matrix
- About Page: Explains system methodology and NAFLD context

6.4 Website Demonstration

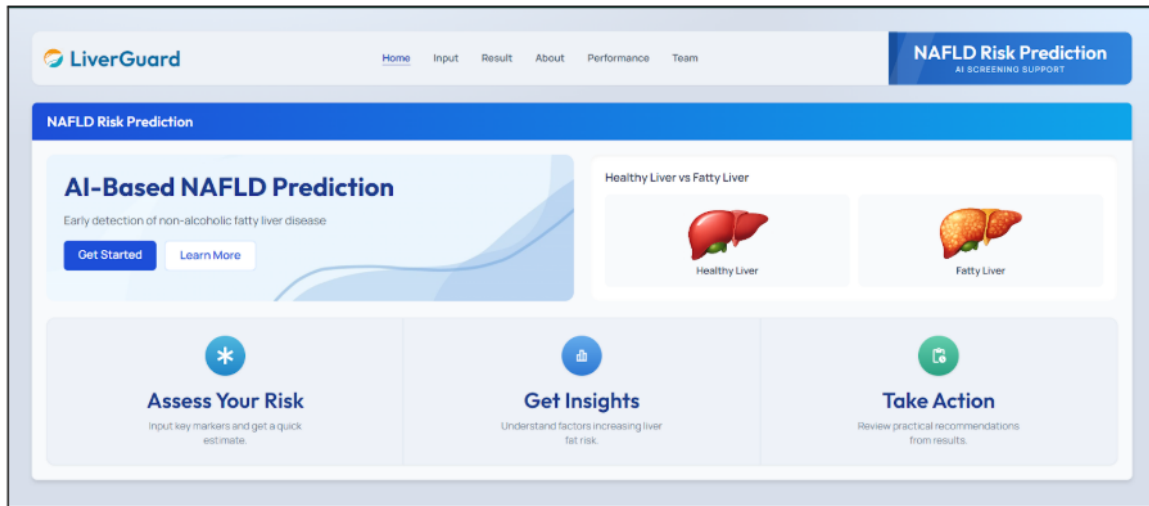


Figure 6.2: LiverGuard Home Page Interface

The LiverGuard application homepage is shown in Figure 6.2. It introduces the AI-powered NAFLD risk calculator through a well-organised structure with “Get Started” and “Learn More” buttons. The page features comparisons of healthy and fatty liver images and a menu bar that contains links to Input, Result, About, Performance and Team pages. The “Get Started” button links to the prediction input form page.

Figure 6.3 shows the input page in its initial state. It contains a structured form for collecting patient clinical and laboratory data such as age, height, weight, BMI, FBS, ALT, AST, lipid profile values, and diabetes status.

All required fields are enforced through validation, preventing incomplete submissions. The “Predict” button is used to submit the data once all fields are filled.

Figure 6.4 shows the data input page after user data entry. An input data preview panel for model request shows the structured input values of BMI, FBS, ALT, AST, LDL, HDL, Triglycerides, Cholesterol level and diabetes status.

This helps format the data properly and ensures the input is correct before it is transmitted to the model for prediction.

Figure 6.5 shows the result page after submitting the data and inference. The result shows the patient is at a Moderate Risk with a 52% probability.

Prediction Input

Enter Patient Details

These values will be sent to the backend prediction model.

Age *

Enter age

Weights *

Enter weights

Height *

Feets

Enter feets

Inches

Enter inches

BMI *

Enter bmi

FBS *

Enter fbs

ALT *

Enter alt

AST *

Enter ast

LDL *

Enter ldl

HDL *

Enter hdl

Triglycerides *

Enter triglycerides

Cholesterol *

Enter cholesterol

Diabetes status *

Select Diabetes Status

Please fill all required fields before prediction.

Predict

Figure 6.3: Prediction Input Form

Prediction Input

Enter Patient Details

These values will be sent to the backend prediction model.

Age *	Weights *
26	87
Height *	
Feets	Inches
6	0
BMI *	FBS *
26.01	99
ALT *	AST *
85	50
LDL *	HDL *
150	28
Triglycerides *	Cholesterol *
225	209
Diabetes status *	
0 = No	

Predict

Model Request Preview

Live form data preview (what will be sent to backend)

```
{
  "age": "26",
  "feet": "6",
  "inches": "0",
  "weight": "87",
  "bmi": "26.01",
  "fbs": "99",
  "alt": "85",
  "ast": "50",
  "ldl": "150",
  "hdl": "28",
  "triglycerides": "225",
  "cholesterol": "209",
  "diabetes_status": "0"
}
```

This shows real-time data being prepared for prediction.

Figure 6.4: Prediction Input Form with Data and Model Request Preview

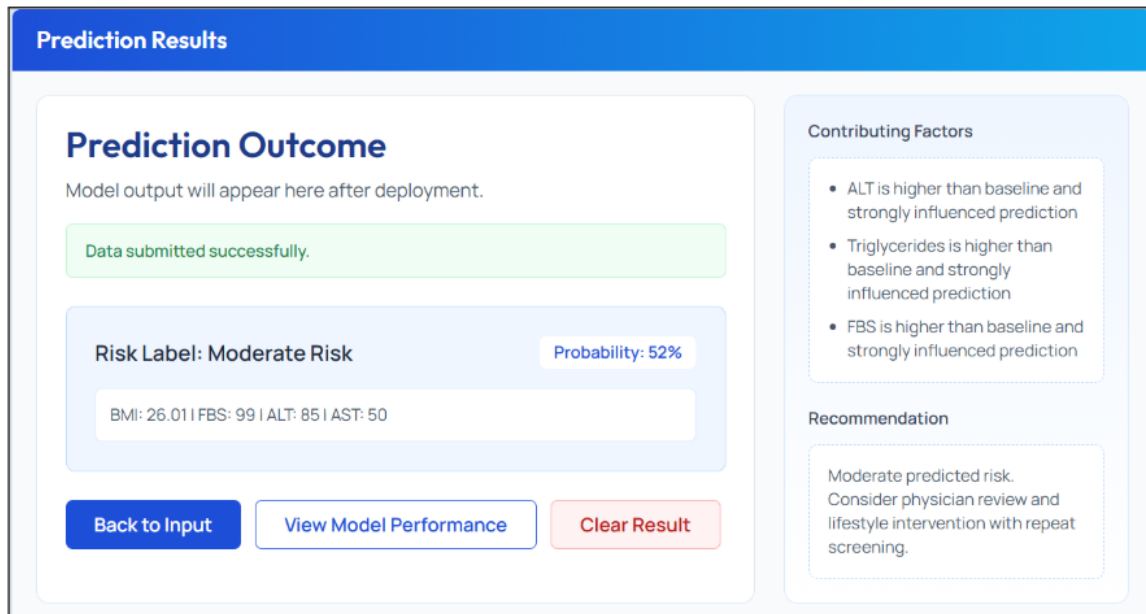


Figure 6.5: Prediction Result Page with Risk Output

It also shows the important input values like BMI: 26.01, FBS: 99, ALT: 85, and AST: 50. The factors that contributed are shown, with high ALT, triglycerides and FBS as major factors.

The recommendation section recommends consulting a doctor and changing lifestyle habits with a follow-up test. The user can choose to go back to enter data, view the performance, or clear the screen.

The top part of the performance page in Figure 6.6 displays assessment metrics with Accuracy: 90.93%, Precision: 85.51%, Recall: 98.57%, and F1-Score: 91.58%. The confusion matrix shows classification results including 2761 NAFLD cases correctly predicted, 2333 non-NAFLD cases correctly predicted and some false positives and negatives. Top influential features are also shown including Height (0.469), ALT (0.124), AST (0.081), FBS (0.064), and BMI (0.054).

The bottom part of the performance page is shown in Figure 6.7. It shows ensemble model signals with weights like Random Forest (14.2070), XGBoost (2.7911), and Logistic Regression (1.3646), which show how much they helped make the final prediction. The page also shows all the input features that the model used, such as age, height, weight, BMI, biochemical markers, and whether or not the person has diabetes. The system also gives information about the model version (v1.0.0),

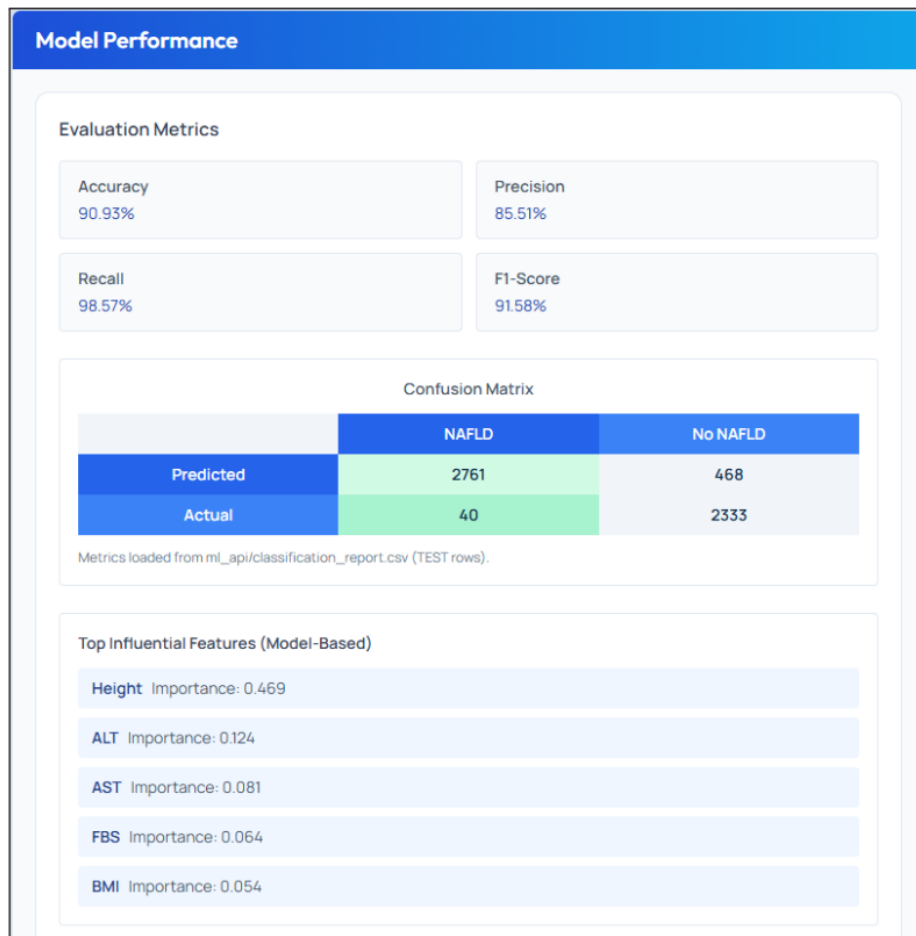


Figure 6.6: Model Performance Page – Evaluation Metrics and Confusion Matrix

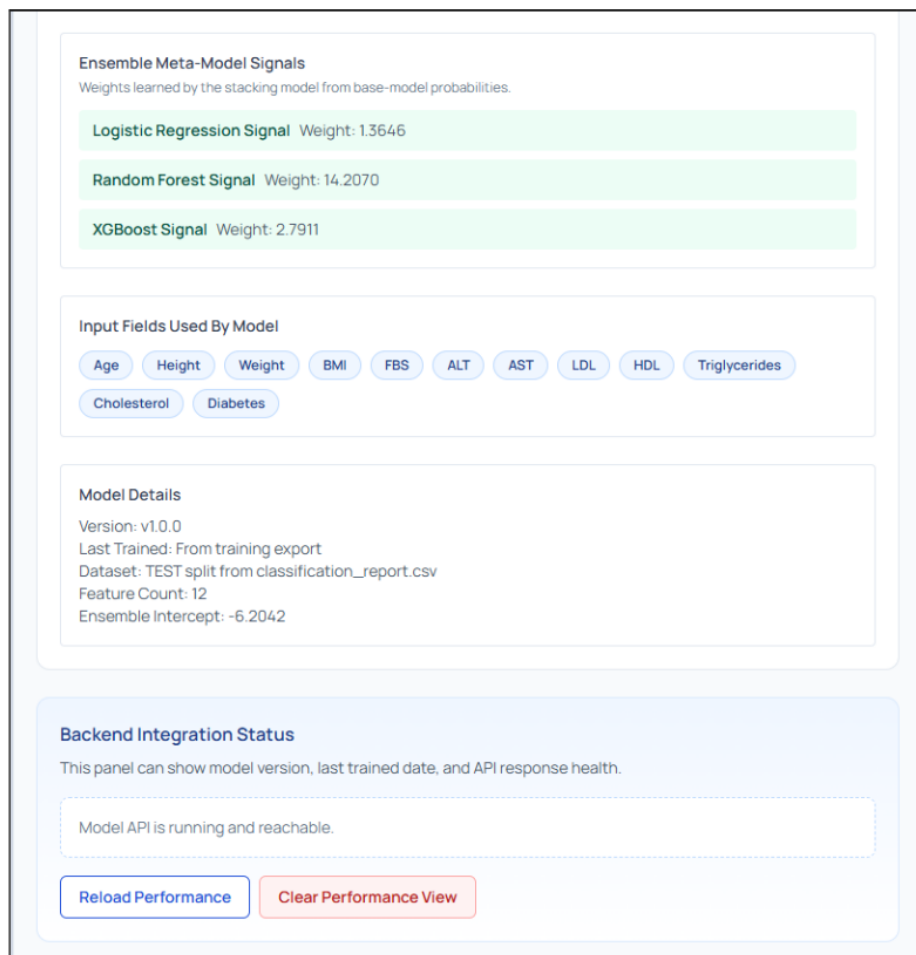


Figure 6.7: Model Performance Page – Ensemble Signals and System Details

the source of the dataset, the number of features (12), and the ensemble intercept (-6.2042). A backend status panel checks that the model API is working and can be reached.



Figure 6.8: About Page – Project Overview and Explainability Components

The top part of the About page for the LiverGuard app can be seen in Figure 6.8. It talks about the project background and how NAFLD is becoming more of a problem, which is why early, non-invasive screening is so important.

The page explains how the system uses routine clinical data to estimate risk and how it combines public datasets with local clinical data. It also talks about important ideas like data fusion with SoftImpute and Explainable AI (SHAP-based interpretation), which gives both global feature importance and patient-level reasoning for predictions.

The lower part of the About page is shown in Figure 6.9. It defines the primary goal of providing scalable and non-invasive NAFLD risk assessment system on the basis of standard medical indicators.

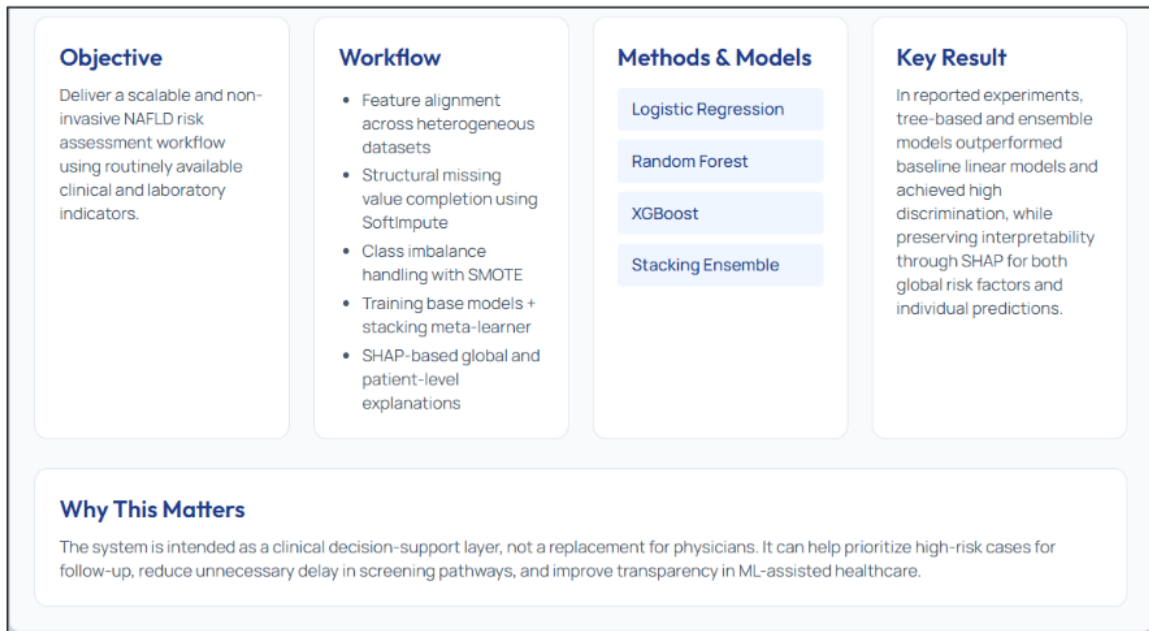


Figure 6.9: About Page – Objectives, Workflow, and Model Details

The steps are outlined in a single workflow, with feature alignment, missing value management with SoftImpute, class imbalance management with SMOTE, and training of base models on a stacking ensemble.

Models used, such as Logistic Regression, Random Forest, XGBoost, and Stacking Ensemble, are also mentioned on the page, and the important findings, including performance improvement and explainability with SHAP, are indicated. An explanation of the use of the system as a clinical decision-support tool is provided in a section titled Why This Matters.

Figure 6.10 displays the Team page of LiverGuard application. It includes the details of the project supervisor and student team that are working on the project.

The name and affiliation of the supervisor appear on the top section and the profile of the individual team members has their name and student ID. This page gives the appropriate recognition of the contributions and is transparent about the ownership of the project.

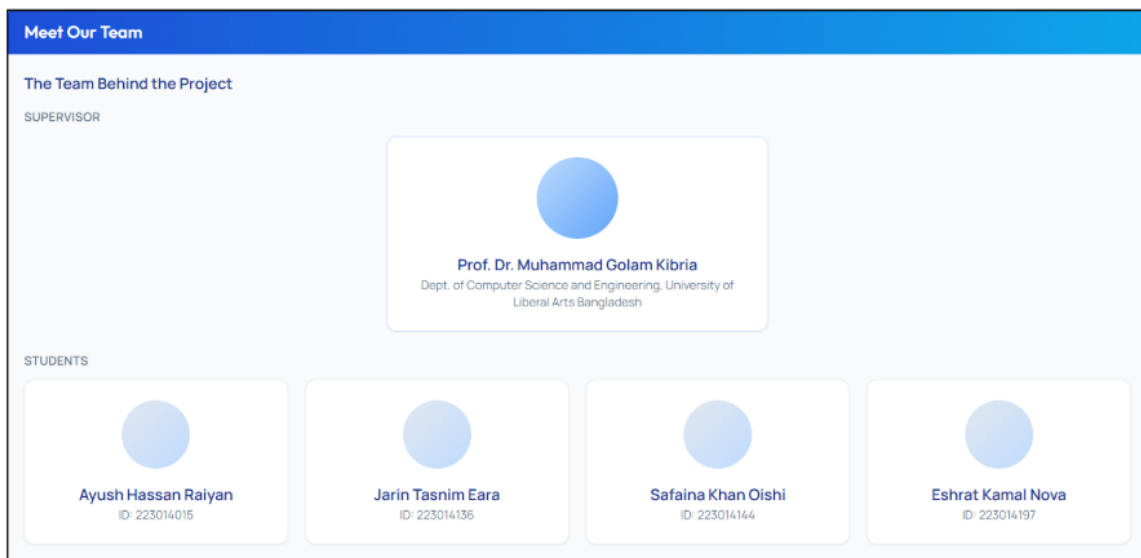


Figure 6.10: Team Page - Project Members and Supervisor

Chapter 7

Conclusions

7.1 Social, Legal, Ethical, and Environmental Issues

7.1.1 Social Issues

In Bangladesh, where access to advanced diagnostic tools is limited, this system can help increase awareness and enable early screening. It supports clinicians by providing fast predictions and explanations, which can improve decision-making and reduce the burden on healthcare systems. However, the system is designed to assist, not replace—medical professionals.

7.1.2 Legal Issues

The system uses publicly available datasets and anonymized clinical data, ensuring that no personally identifiable information is exposed. The design follows basic data protection principles, including safe handling and processing of patient data. For real-world deployment, the system would need to comply with healthcare data regulations and policies, such as data privacy laws and institutional guidelines. Proper authorization and validation would be required before clinical use.

7.1.3 Ethical Issues

Explainability is provided by SHAP to ensure transparency and fairness of the system ethically. This will enable clinicians to know why predictions are made. It also minimizes the black-box characteristics of machine learning models. To minimize bias, several datasets were concatenated and then balanced with SMOTE.

But there is a possibility of bias as datasets might not be a full representation of the entire population. Thus, medical professionals should always revise their predictions and finalize decisions.

7.1.4 Environmental Issues

This project has minimal environmental impact because it is a software-based solution. It reduces the need for invasive procedures like liver biopsies, which require medical resources and materials. The system runs on standard computing infrastructure which consumes relatively low energy. Overall, the solution supports a more sustainable and efficient healthcare approach.

7.2 Brief Summary

An explainable machine learning system based on fusion was developed in order to predict the risk of NAFLD using public datasets as well as real-world Bangladeshi clinical data. The system is able to solve several critical issues including the heterogeneity of data and missing values with the help of matrix decomposition (SoftImpute). It also enhances the performance of the model with the help of SMOTE-based balancing and ensemble learning.

Various models were trained and tested with the stacking ensemble model performing the best (more than 99% accuracy). The explainability provided by SHAP allows global and patient-level interpretation which makes the system more transparent and useful in clinical practice.

Moreover, a web-based application was designed to deliver real-time predictions and explanations so that the system can be used in the real world. In general, the project has shown a scalable and reliable method of early NAFLD risk assessment.

7.3 Group Members Contribution

The following section outlines the specific roles and technical contributions of each team member toward the successful completion of this research project:

- **Ayush Hassan Raiyan:** Conducted the literature review and coordinated data collection. Led the implementation of the XGBoost and Stacking ensemble

bles. Responsible for dataset generation via Matrix Decomposition (merged) and the development of the Logistic Regression (LR) model for the website interface.

- **Jarin Tasnim Eara:** Contributed to the literature review and data collection phases. Implemented the Logistic Regression (LR) baseline and developed the Zero-padding merging technique. Led the development of the XGBoost website model and the final Website Model Ensemble.
- **Safaina Khan Oishi:** Performed the literature review and data collection. Specialized in model interpretability by implementing SHAP for the stacking ensemble model. Deployed the core website models and integrated SHAP explanations for the XGBoost website model and the overall stacking architecture.
- **Eshrat Kamal Nova:** Contributed to the literature review and data collection. Developed the Random Forest (RF) models for both the primary research and the website interface. Led the frontend development and backend connections of the web application.

7.4 Future works

In the future, this work can be extended by incorporating larger and more diverse clinical datasets to further improve the model's performance and reliability across different populations. The system can be validated using real hospital data and clinical trials to ensure its effectiveness in real-world healthcare environments. The web application can be enhanced by adding real-time database integration, secure user authentication, and improved user interfaces for better usability. Additionally, exploring advanced techniques such as deep learning and multi-modal data integration, including imaging and clinical data, can further improve prediction accuracy. The explainability component can also be enhanced with more interactive visualization tools to support better clinical understanding. Finally, deploying the system as a mobile application can increase accessibility and enable wider adoption, especially in resource-limited settings.

References

- [1] Y. X. Liu *et al.*, “Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study,” *Hepatobiliary & Pancreatic Diseases International*, vol. 20, no. 5, pp. 409–415, 2021.
- [2] L. Liu *et al.*, “Automated machine learning models for nonalcoholic fatty liver disease assessed by controlled attenuation parameter from the NHANES 2017–2020,” *Digital Health*, vol. 10, pp. 1–19, 2024.
- [3] H. Y. Peng *et al.*, “Development and validation of machine learning models for nonalcoholic fatty liver disease,” *Hepatobiliary & Pancreatic Diseases International*, vol. 22, no. 6, pp. 615–621, 2023.
- [4] A. Alizargar *et al.*, “Precision non-alcoholic fatty liver disease (NAFLD) diagnosis: Leveraging ensemble machine learning and gender insights for cost-effective detection,” *Bioengineering*, vol. 11, no. 6, p. 600, 2024.
- [5] S. Rokoni *et al.*, “Detection of nonalcoholic fatty liver disease using deep learning algorithms,” *Journal of Computer Science and Technology Studies*, vol. 5, no. 4, pp. 150–159, 2023.
- [6] B. B. Munukuntla and M. S. Yalawar, “Machine learning algorithms-based non-alcoholic fatty liver disease prediction,” *International Journal of Recent Technology and Engineering*, vol. 12, no. 3, pp. 43–46, 2023.
- [7] F. M. Alshagathrh and M. S. Househ, “Artificial intelligence for detecting and quantifying fatty liver in ultrasound images: A systematic review,” *Bioengineering*, vol. 9, no. 12, p. 748, 2022.
- [8] B. Yang, H. Lu, and Y. Ran, “Advancing non-alcoholic fatty liver disease prediction: A comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation,” *Frontiers in Endocrinology*, vol. 15, p. 1450317, 2024.
- [9] L. R. Guarneros-Nolasco *et al.*, “Early identification of risk factors in non-alcoholic fatty liver disease (NAFLD) using machine learning,” *Mathematics*, vol. 11, no. 13, p. 3026, 2023.

- [10] J. L. Wang *et al.*, “Non-invasive diagnosis of non-alcoholic fatty liver disease: Current status and future perspective,” *Heliyon*, vol. 10, p. e27325, 2024.
- [11] P. Decharatanachart *et al.*, “Application of artificial intelligence in non-alcoholic fatty liver disease and liver fibrosis: a systematic review and meta-analysis,” *Therapeutic Advances in Gastroenterology*, vol. 14, 2021.
- [12] P. Sorino *et al.*, “Development and validation of a neural network for NAFLD diagnosis,” *Scientific Reports*, vol. 11, no. 1, p. 20240, 2021.
- [13] J. R. Zheng, Z. L. Wang, and B. Feng, “Machine learning-based mortality prediction models for non-alcoholic fatty liver disease in the general United States population,” *medRxiv*, 2024.
- [14] L. Xiao *et al.*, “Development and validation of machine learning-based marker for early detection and prognosis stratification of nonalcoholic fatty liver disease,” *Advanced Science*, p. e10527, 2025.
- [15] O. Kulthe *et al.*, “Classification of NAFLD using ML algorithms,” in *Proc. Int. Conf. on Advances in Modern Age Technologies for Health and Engineering Science*, 2025, pp. 1–4.
- [16] B. Njei *et al.*, “An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis,” *Scientific Reports*, vol. 14, no. 1, p. 8589, 2024.
- [17] D. Srivastava *et al.*, “Stacking ensemble learning approach for non-alcoholic fatty liver disease identification: Leveraging explainable machine learning for enhanced prediction models,” *Frontiers in Biomedical Technologies*, vol. 12, no. 3, 2026.
- [18] S. Savas, “Explainable artificial intelligence for diagnosis and staging of liver cirrhosis using stacked ensemble and multi-task learning,” *Diagnostics*, vol. 15, no. 9, p. 1177, 2025.
- [19] J. Płudowski and J. Mulawka, “Application of artificial intelligence in support of NAFLD diagnosis,” *Applied Sciences*, vol. 14, no. 22, p. 10237, 2024.
- [20] Y. Chen *et al.*, “Multi-modal learning-based prediction for disease,” *arXiv preprint arXiv:2307.09823*, 2023.
- [21] Y. Chen *et al.*, “Multimodal learning-based prediction for nonalcoholic fatty liver disease,” *Machine Intelligence Research*, pp. 1–17, 2025.
- [22] Y. Hong *et al.*, “Machine learning prediction of metabolic dysfunction-associated fatty liver disease risk in American adults using body composition: explainable analysis based on SHapley Additive exPlanations,” *Frontiers in Nutrition*, vol. 12, p. 1616229, 2025.

- [23] S. Bao *et al.*, “Predicting the risk of lean non-alcoholic fatty liver disease based on interpretable machine models in a Chinese T2DM population,” *Frontiers in Endocrinology*, vol. 16, p. 1626203, 2025.
- [24] N. Babakov *et al.*, “On the role of explanations in machine learning prediction of nonalcoholic steatohepatitis,” *Progress in Artificial Intelligence*, vol. 14, p. 387, 2025.
- [25] W. Ji *et al.*, “A machine learning–based framework to identify and classify Non-alcoholic Fatty Liver Disease in a Large-Scale Population,” *Frontiers in Public Health*, vol. 10, p. 846118, 2022.
- [26] G. Zhu *et al.*, “Explainable machine learning model for predicting the risk of significant liver fibrosis in patients with diabetic retinopathy,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 332, 2024.
- [27] G. Arya *et al.*, “Explainable AI for enhanced interpretation of liver cirrhosis biomarkers,” *IEEE Access*, vol. 11, pp. 123729–123741, 2023.
- [28] J. Xi *et al.*, “Life’s Crucial 9 and NAFLD using SHAP-interpreted machine learning predictions,” *Scientific Reports*, vol. 15, no. 1, p. 9384, 2025.
- [29] T. C. Yip *et al.*, “Non-invasive biomarkers for liver inflammation in non-alcoholic fatty liver disease: present and future,” *Clinical and Molecular Hepatology*, vol. 29, Suppl., pp. S171–S183, 2023.
- [30] M. Fan *et al.*, “A deep matrix completion method for imputing missing histological data,” *Medical Physics*, 2021.
- [31] M. Huang *et al.*, “Missing value imputation with low-rank matrix completion in single-cell RNA-seq,” *Frontiers in Genetics*, vol. 13, p. 952649, 2022.
- [32] M. Du *et al.*, “Meta-learning-based inductive logistic matrix completion,” *Journal of Cheminformatics*, vol. 16, no. 1, p. 44, 2024.
- [33] F. R. Zhang *et al.*, “Non-linear matrix completion,” *Journal of Data Analysis and Information Processing*, vol. 12, pp. 115–137, 2024.
- [34] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *JMLR*, vol. 11, pp. 2287–2322, 2010.
- [35] B. Ramana and N. Venkateswarlu, “ILPD (Indian Liver Patient Dataset),” UCI Machine Learning Repository, 2022.
- [36] Utkarshx27, “Non-Alcohol Fatty Liver Disease Dataset,” Kaggle, 2022.
- [37] Sunilsah905, “Non-Alcoholic Fatty Liver Disease Dataset,” Kaggle, 2021.
- [38] JesusDeLeon19, “Non-Alcoholic Fatty Liver Dataset,” Kaggle, 2021.