Reinforcing synthetic data of patients suffering from liver cirrhosis

By

Anai Sunny

Submitted to
The University of Roehampton

In partial fulfilment of the requirements
for the degree of

Master of Science

in

Data Science

DECLARATION

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Anai Sunny

06/05/2025

ACKNOWLEDGEMENTS

I express my sincere gratitude to all who contributed to the successful completion of this liver disease prediction project. I am deeply thankful to my academic supervisor Shekoufeh Rahimi for their invaluable guidance, constructive feedback, and encouragement throughout the research process, which helped shape the direction of this study. I also extend my appreciation to my peers and colleagues who provided insightful discussions and technical support, particularly in navigating the complexities of machine learning algorithms and data preprocessing techniques. I am grateful to the University of California, Irvine (UCI) Machine Learning Repository for providing access to the Indian Liver Patient Dataset (ILPD), which served as the foundation for this research. Finally, I am indebted to my family and friends for their unwavering support and encouragement, which motivated me to overcome challenges and complete this project successfully.

ABSTRACT

Liver disease remains a significant global health challenge, necessitating early and accurate prediction to improve patient outcomes.This study develops a robust machine learning framework to predict liver disease using the Indian Liver Patient Dataset (ILPD), which comprises 583 patient records with 10 biochemical and demographic features. The primary objective is to identify key predictive markers and build an effective classification model capable of addressing class imbalance, a common issue in medical datasets.

The methodology begins with exploratory data analysis (EDA) to uncover patterns and relationships within the data. EDA revealed that log-transformed biochemical markers, such as log_Direct_Bilirubin (correlation: 0.31), log_Total_Bilirubin (0.31), log_Aspartate_Aminotransferase (0.30), and log_Alamine_Aminotransferase (0.28), were the strongest predictors of liver disease, while protein-related features like Total_Proteins (-0.04) showed limited discriminatory power. Visualizations, including boxplots and pair plots, confirmed higher bilirubin and enzyme levels in diseased patients, alongside a gender imbalance (75% male) and age distribution peaking at 40–60 years. Preprocessing involved imputing missing values in Albumin_and_Globulin_Ratio with the mean (0.947), removing duplicates, and encoding categorical variables (Gender: Male=1, Female=0). Feature engineering enhanced predictive power by applying log transformations to skewed biochemical features, binning Age into four groups, and creating derived features like Bilirubin_Ratio and Enzyme_Sum, with outliers capped at the 5th and 95th percentiles.

The dataset's class imbalance (71.4% diseased, 28.6% non-diseased) was addressed using two balancing techniques: SMOTETomek, which oversamples the minority class and removes overlapping majority samples, and Variational Autoencoder (VAE), which generates synthetic minority samples. Five machine learning algorithms—Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Neural Network—were trained under three conditions: original imbalanced data, SMOTETomek-balanced data, and VAE-balanced data. Performance evaluation utilized accuracy, F1 macro score, class-specific recalls, and ROC curves. Without balancing, models were biased toward the majority class, with SVM and Neural Network achieving 1.00 and 0.99 recall for Class 1 (disease) but 0.00 for Class 0 (no disease). SMOTETomek significantly improved performance, with the Neural Network achieving the best results: 75% accuracy, 0.73 F1 macro, recalls of 0.79 (Class 0) and 0.74 (Class

1), and an AUC of 0.76. VAE balancing maintained accuracy but yielded a lower F1 macro (0.65 for Neural Network). ROC curve analysis further confirmed SMOTETomek's superiority, with the Neural Network's AUC peaking at 0.76, reflecting balanced discriminative abilityThese findings contribute to the development of reliable predictive models for early liver disease detection, offering practical insights for healthcare applications and aligning with prior research on machine learning in medical diagnostics.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. Chapter 1 INTRODUCTION

Liver disease represents a critical public health challenge worldwide, contributing significantly to morbidity and mortality rates. Conditions such as cirrhosis, hepatitis, and fatty liver disease often progress silently, with symptoms manifesting only at advanced stages, making early detection paramount for effective intervention and improved patient outcomes. According to global health statistics, liver-related diseases account for over 2 million deaths annually, a Figureure exacerbated by factors like alcohol consumption, viral infections, obesity, and metabolic disorders. Traditional diagnostic methods, such as liver biopsies and imaging techniques, are often invasive, costly, and inaccessible in resource-limited settings, underscoring the need for non-invasive, cost-effective alternatives. Advances in machine learning (ML) have opened new avenues for medical diagnostics, offering the potential to analyze complex clinical datasets and identify patterns that may elude conventional approaches. By leveraging biochemical and demographic data, ML models can predict the likelihood of liver disease with high accuracy, enabling timely interventions and personalized treatment plans.

The application of machine learning in liver disease prediction has gained traction in recent years, driven by the availability of structured medical datasets and the ability of ML algorithms to handle high-dimensional data. However, a significant challenge in this domain is the prevalence of class imbalance, where the number of diseased patients often far exceeds non-diseased ones, leading to biased models that perform poorly on minority classes. This imbalance can skew predictions, resulting in high false-negative rates for non-diseased cases, which is particularly problematic in medical contexts where missing a diagnosis can have severe consequences. Addressing this issue requires robust data preprocessing and balancing techniques, such as synthetic data generation and resampling methods, to ensure models are equitable and reliable across all classes. Furthermore, the identification of key predictive features—such as biochemical markers like bilirubin and liver enzymes—is critical for building interpretable and clinically relevant models, as these features directly correlate with liver function and disease progression.

This project focuses on developing a machine learning framework to predict liver disease using the Indian Liver Patient Dataset (ILPD), a publicly available dataset comprising 583 patient records with 10 features, including age, gender, and biochemical markers like total bilirubin, direct bilirubin, alkaline phosphatase, and liver enzymes (ALT, AST).

The dataset, collected from patients in North East Andhra Pradesh, India, exhibits a class imbalance with 71.4% of records classified as diseased and 28.6% as non-diseased, posing a challenge for model training. The primary objective of this study is to design a predictive model that accurately classifies liver disease while addressing class imbalance through advanced techniques like SMOTETomek and Variational Autoencoder (VAE)-based synthetic data generation.

The methodology encompasses exploratory data analysis (EDA) to uncover feature relationships, preprocessing to handle missing values and normalize distributions, feature engineering to enhance predictive power, and the evaluation of five machine learning algorithms—Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Neural Network—under three conditions: original imbalanced data, SMOTETomek-balanced data, and VAE-balanced data. Performance is assessed using accuracy, F1 macro score, class-specific recalls, and ROC curve analysis to ensure balanced classification. By identifying the most predictive biochemical markers and optimizing model performance, this study aims to contribute to the development of reliable, non-invasive tools for early liver disease detection, ultimately supporting healthcare professionals in delivering timely and effective care.

1.1 PROBLEM DESCRIPTION, CONTEXT AND MOTIVATION

a. Problem Description

Liver disease prediction poses a significant challenge in medical diagnostics due to the complexity of identifying early-stage conditions using non-invasive methods.

The primary problem addressed in this project is the accurate classification of patients with liver disease versus those without, using clinical data from the Indian Liver Patient Dataset (ILPD).

The ILPD contains 583 patient records, each with 10 features, including biochemical markers (e.g., total bilirubin, direct bilirubin, alkaline phosphatase, ALT, AST) and demographic variables (age, gender). A key issue is the dataset's class imbalance, with 71.4% of records labeled as diseased (416 patients) and 28.6% as non-diseased (167 patients), which can bias machine learning models toward the majority class, leading to poor detection of non-diseased cases. Additionally, the dataset exhibits feature skewness (e.g., bilirubin and enzyme levels are right-skewed) and missing values in the Albumin_and_Globulin_Ratio feature, requiring careful preprocessing to ensure model reliability. The goal is to develop a machine learning framework that leverages these features to predict liver disease accurately, addresses class imbalance through techniques like SMOTETomek and Variational Autoencoder (VAE), and evaluates performance using metrics such as accuracy, F1 macro score, class-specific recalls, and ROC curves across five algorithms: Random Forest, Logistic Regression, Decision Tree, SVM, and Neural Network.

b. Context

Liver disease is a pressing global health concern, contributing to over 2 million deaths annually, with conditions like cirrhosis, hepatitis, and non-alcoholic fatty liver disease (NAFLD) on the rise due to factors such as alcohol consumption, viral infections, and metabolic syndromes.

Traditional diagnostic methods, such as liver biopsies and imaging, are often invasive, expensive, and impractical for widespread use, especially in rural or underserved regions.

Non-invasive alternatives, such as blood-based biomarkers, offer a promising solution, but their interpretation requires sophisticated analysis to detect subtle patterns indicative of early-stage disease. Machine learning has emerged as a powerful tool in this context, capable of analyzing complex clinical datasets to identify predictive patterns and support early diagnosis. Prior studies have demonstrated the potential of ML in liver disease prediction, yet challenges like class imbalance and feature selection remain underexplored, particularly in the context of the ILPD, which reflects a specific demographic and clinical profile.

c. Motivation

The motivation for this project stems from the urgent need to improve early detection of liver disease, which can significantly enhance patient outcomes through timely intervention. The high mortality rates associated with late-stage liver disease underscore the importance of developing accessible, non-invasive diagnostic tools that can be deployed in diverse healthcare settings, including resource-constrained environments like those in India. Machine learning offers a unique opportunity to address this need by leveraging readily available clinical data, such as the biochemical markers in the ILPD, to build predictive models that are both accurate and interpretable.

The class imbalance in the ILPD presents a specific challenge that, if addressed effectively, can lead to more equitable models with better generalization across both diseased and non-diseased cases, a critical requirement in medical applications where false negatives can have severe consequences. Additionally, the project is motivated by the potential to contribute to the growing body of research on ML in healthcare, particularly by exploring advanced balancing techniques like SMOTETomek and VAE, and comparing their impact on model performance. By identifying the most predictive features and optimizing classification performance, this study aims to provide actionable insights for clinicians and pave the way for scalable, data-driven solutions in liver disease management.

1.2 OBJECTIVES

The following objectives outline the key technical goals of the liver disease prediction project, focusing on the implementation aspects using the Indian Liver Patient Dataset (ILPD). These objectives aim to ensure a robust, accurate, and balanced classification model through systematic data processing, feature enhancement, and model evaluation.

- ➢ Conduct Comprehensive Data Preprocessing to Ensure Data Quality: Implement a series of preprocessing steps to prepare the ILPD for modeling, including imputing missing values in the Albumin_and_Globulin_Ratio feature using the mean (0.947), removing duplicate entries to reduce the dataset to 579 unique records, encoding categorical variables (Gender: Male=1, Female=0; Liver_Disease: 1=disease, 0=no disease), and standardizing numerical features using StandardScaler to normalize distributions for machine learning algorithms.

- Enhance Predictive Power Through Targeted Feature Engineering: Apply feature engineering techniques to improve model performance by addressing skewness and creating new features, specifically applying log transformations to right-skewed biochemical features (Total_Bilirubin, Direct_Bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase), binning the Age feature into four categorical groups (0–18, 18–35, 35–50, 50–90 years), deriving new features such as Bilirubin_Ratio (Direct/Total Bilirubin) and Enzyme_Sum (ALT + AST), and capping outliers at the 5th and 95th percentiles to mitigate their impact on model training.
- Address Class Imbalance Using Advanced Balancing Techniques: Implement two balancing strategies to tackle the ILPD's class imbalance (71.4% diseased, 28.6% non-diseased), first by applying SMOTETomek to oversample the minority class (non-diseased) and remove overlapping majority class samples, and second by using a Variational Autoencoder (VAE) with an architecture consisting of an encoder (input → 16-unit hidden layer → 8-unit latent space) and decoder to generate synthetic minority class samples, ensuring balanced training data for fair model performance.
- Train and Optimize Multiple Machine Learning Models for Classification: Develop and train five machine learning models—Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Neural Network—under three conditions: original imbalanced data, SMOTETomek-balanced data, and VAE-balanced data, using a train-test split (80:20 ratio, stratified by Liver_Disease), optimizing hyperparameters (e.g., Neural Network with 64-32-1 architecture, dropout layers of 0.3 and 0.2, and early stopping with patience=30), and ensuring reproducibility with a random state of 42.
- Evaluate Model Performance Using Comprehensive Metrics and Visualizations: Assess the performance of all models using a suite of metrics including accuracy, F1 macro score, class-specific recalls (for Class 0: no disease, and Class 1: disease), and Area Under the ROC Curve (AUC), generating ROC curves for each algorithm under the three balancing conditions to visualize discriminative ability, with a focus on achieving balanced performance across both classes, particularly improving minority class

detection as evidenced by the Neural Network's SMOTETomek results (75% accuracy, 0.73 F1 macro, AUC 0.76).

1.3 METHODOLOGY

This section outlines the methodology for predicting liver disease using the Indian Liver Patient Dataset (ILPD), sourced from Kaggle and originally provided by the UCI Machine Learning Repository. The methodology focuses on the functional steps undertaken to preprocess the data, engineer features, analyze correlations, visualize insights, balance classes, train models, and evaluate performance.

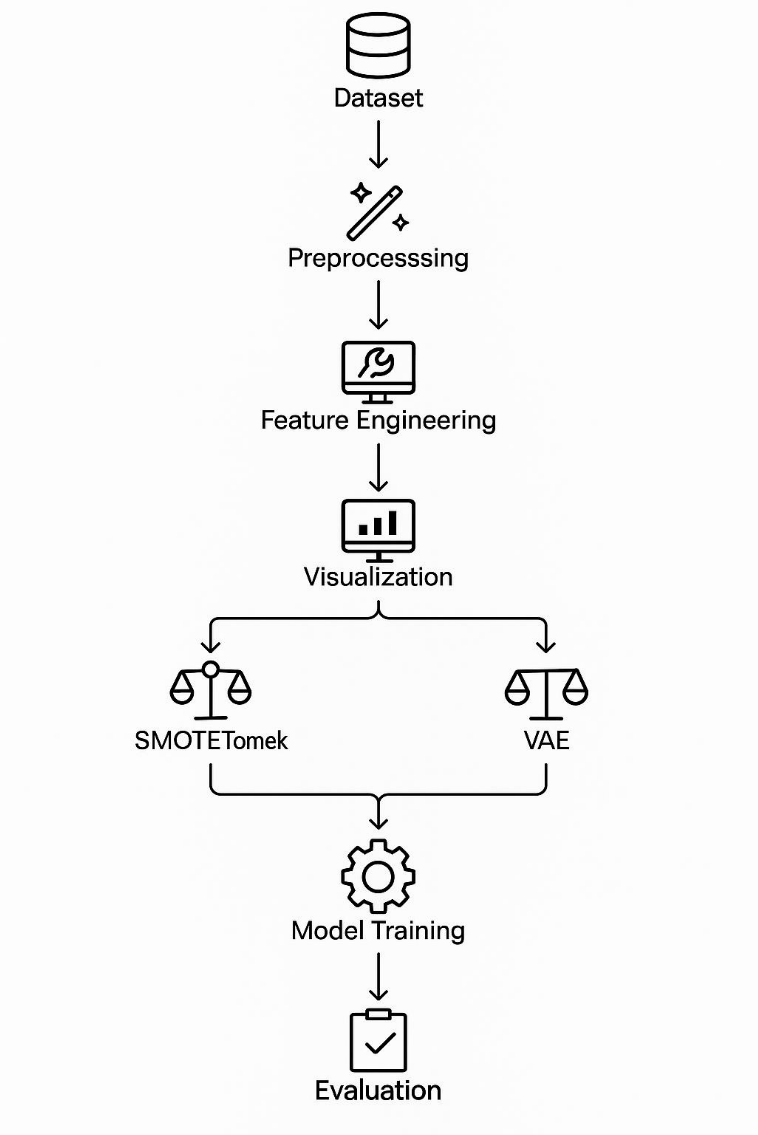


Figure 1 Architecture Diagram

a. Dataset

The ILPD dataset was obtained from Kaggle, a platform hosting publicly available datasets for machine learning research, originally contributed by the UCI Machine Learning Repository. It comprises 583 patient records from North East Andhra Pradesh, India, with the following characteristics:

- Class Distribution: 416 patients with liver disease (71.4%) and 167 without (28.6%), indicating a class imbalance.
- Gender Distribution: 441 males (75.6%) and 142 females (24.4%).
- Age Range: 4 to 90 years, with patients over 89 recorded as 90.
- Attributes: 11 features, including age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, albumin-globulin ratio, and the target variable indicating liver disease status.

Table 1 Description of Dataset Columns

Column Name	Description
Age	Age of the patient in years
Gender	Gender of the patient (Male or Female)
Total_Bilirubin	Total bilirubin level in the blood (mg/dL)
Direct_Bilirubin	Direct (conjugated) bilirubin level in the blood (mg/dL)
Alkaline_Phosphotase	Level of alkaline phosphatase enzyme (IU/L)
Alamine_Aminotransferase	Level of ALT (SGPT) enzyme in the blood (IU/L)
Aspartate_Aminotransferase	Level of AST (SGOT) enzyme in the blood (IU/L)
Total_Protiens	Total protein content in blood (g/dL)
Albumin	Albumin level in the blood (g/dL)
Albumin_and_Globulin_Ratio	Ratio of albumin to globulin in the blood
Dataset	Target variable: 1 = Liver disease, 2 = No liver disease

The dataset was selected to develop machine learning models for liver disease prediction, aiming to reduce diagnostic burden by leveraging biochemical markers and demographic data to identify disease patterns.

b. Preprocessing

Preprocessing was conducted to ensure the dataset was suitable for analysis and modeling. The dataset was loaded directly from a GitHub URL, confirming its structure with 583 records and 11 columns. The target variable was converted to a binary format (1 for liver disease, 0 for no disease) to align with binary classification requirements. Missing values, totaling four entries in the albumin-globulin ratio feature, were imputed using the feature’s mean to maintain data completeness. Gender was encoded numerically (Male as 1, Female as 0) to enable its use in computational models. Duplicate records were removed, reducing the dataset to 579 unique entries, ensuring data quality. Initial exploration included generating statistical summaries to understand feature distributions, revealing:

- Age: Mean 44.75 years, ranging from 4 to 90.
- Gender: Predominantly male (75.6%).
- Biochemical Markers: Significant skewness in features like total bilirubitrellon (mean: 3.30, max: 75.0) and aspartate aminotransferase (mean: 109.91, max: 4929).

These steps ensured a clean dataset ready for feature engineering and analysis.

c. Feature Engineering

Feature engineering was performed to enhance the dataset’s predictive capabilities by addressing skewness and creating new features. The process involved several functional transformations:

- Normalization of Skewed Features: Biochemical markers such as total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, and aspartate aminotransferase exhibited significant right-skewness. A logarithmic transformation was applied to these features (adding 1 to handle zero values), creating normalized versions to improve their distribution for modeling.
- Age Categorization: Age was discretized into four categories (0–18, 18–35, 35–50, 50–90 years) and encoded as integers (0 to 3) to capture age-related patterns in liver disease prevalence.
- Derived Features: Two new features were created to capture additional relationships: a bilirubin ratio (direct bilirubin divided by total bilirubin, with a small epsilon to avoid division by zero) to reflect the proportion of conjugated bilirubin, and an enzyme sum (alamine aminotransferase plus aspartate aminotransferase) to represent combined liver enzyme activity.

- Outlier Mitigation: Extreme values in the original skewed biochemical features were capped at the 5th and 95th percentiles to reduce their impact on model training, ensuring robustness.

These transformations improved the dataset's suitability for predictive modeling by normalizing distributions and introducing meaningful features.

d. Correlation Check with Liver Disease

Correlation analysis was conducted to evaluate the relationship between features and the liver disease target, guiding feature selection. Correlation coefficients were computed and visualized as a bar plot, revealing key insights:

- Strongest Positive Correlations: Log-transformed biochemical features showed the highest correlations with liver disease, including log direct bilirubin (0.31), log total bilirubin (0.31), log aspartate aminotransferase (0.30), and log alamine aminotransferase (0.28).
- Moderate Positive Correlations: Original biochemical features like direct bilirubin (0.29), total bilirubin (0.27), and alkaline phosphatase (0.23) also correlated positively.
- Weak/Negative Correlations: Features like total proteins (-0.04), albumin, and albumin-globulin ratio showed minimal or inverse relationships with liver disease.
- Demographic and Derived Features: Age (0.13), gender (0.08), enzyme sum (0.17), and bilirubin ratio (0.16) exhibited low to moderate correlations.

This analysis highlighted the importance of bilirubin and liver enzyme features as key predictors, with log transformations enhancing their predictive power by normalizing their distributions.

e. Clustered Correlation Heatmap Analysis

A pairwise correlation analysis was performed across all features, visualized as a clustered heatmap to identify inter-feature relationships and potential multicollinearity. The heatmap revealed distinct clusters of correlated features:

- Bilirubin Cluster: Total bilirubin and direct bilirubin showed a high correlation (0.87), with their log-transformed versions at 0.95, indicating strong interdependence.

- Enzyme Cluster: Alamine aminotransferase and aspartate aminotransferase were highly correlated (0.92), with log versions at 0.94, and the enzyme sum feature correlated at 0.96, reflecting their shared diagnostic relevance.
- Protein Cluster: Total proteins, albumin, and albumin-globulin ratio formed a separate cluster with weak correlations to disease markers (0.01–0.04), suggesting limited predictive value.
- Correlation with Liver Disease: Biochemical features showed correlations of 0.2–0.3 with the target, consistent with the prior analysis.

This analysis confirmed the effectiveness of log transformations in preserving feature relationships while addressing skewness, and identified multicollinearity that could impact model performance, informing feature selection strategies.

f. Visualization

Exploratory visualizations were conducted to gain insights into data distributions, feature relationships, and class differences, aiding in model development:

- Age Distribution: A histogram with kernel density estimation showed a bell-shaped distribution peaking at 40–60 years, with a slight right skew (range: 4–90), indicating a middle-aged and older patient population.
- Gender Distribution: A count plot confirmed a gender imbalance, with 441 males and 142 females, reflecting a male-dominated sample.
- Liver Disease by Gender: A grouped bar chart revealed higher disease prevalence in males (310 cases) compared to females (90 cases), highlighting both class and gender imbalances.
- Total Bilirubin vs. Liver Disease: Boxplots indicated higher bilirubin levels in diseased patients (median ~1.5 mg/dL, IQR: 0.8–3.2 mg/dL) compared to non-diseased (median ~0.9 mg/dL, IQR: 0.7–1.2 mg/dL), with more variability in the diseased group.
- Alkaline Phosphatase Distribution: A histogram showed a right-skewed distribution, with most values between 150–250 IU/L, a long tail, and a peak at 700 IU/L, suggesting abnormal values in some patients.

- Liver Enzymes by Liver Disease: Violin and boxplots of alamine aminotransferase and aspartate aminotransferase showed higher medians and wider distributions in diseased patients, with aspartate aminotransferase outliers up to 350 IU/L, underscoring their diagnostic relevance.
- Pair Plots: Features like age, total bilirubin, direct bilirubin, and alkaline phosphatase showed elevated values in diseased patients, confirming their predictive potential.
- Joint Plot (Age vs. Total Bilirubin): Diseased patients exhibited higher bilirubin levels, with a slight skew toward older ages, indicating age-related disease patterns.
- Scatter Plot (AST vs. Total Bilirubin): A regression line revealed a moderate positive correlation, with outliers suggesting severe liver dysfunction in some cases.
- Swarm Plot (Total Proteins): Displayed overlapping distributions across disease states, indicating limited discriminatory power of total proteins.

These visualizations provided critical insights into feature behavior and their association with liver disease, guiding subsequent modeling steps.

g. Normalization & Data Splitting

Data preparation involved standardizing features and splitting the dataset for training and testing. Features were separated into an independent matrix (excluding the target) and a target vector. Standardization was applied to ensure zero mean and unit variance, addressing scale differences across features (e.g., age mean: 44.75, std: 16.19; alkaline phosphatase mean: 290.58, std: 242.94). The dataset was split into 80% training (464 samples) and 20% testing (115 samples) sets, with stratification to preserve the class distribution (71.4% diseased, 28.6% non-diseased) in both sets, and a random state of 42 for reproducibility.

h. SmoteTomek (Data Balance)

To address the class imbalance (71.4% diseased, 28.6% non-diseased), a hybrid resampling technique, SMOTETomek, was applied to the training data. SMOTETomek combines Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for the minority class (non-diseased) and Tomek links to remove overlapping majority class samples, aiming for a balanced training set. This technique was applied with a random state of 42, and its effectiveness was confirmed through count plots, which showed a balanced class distribution post-resampling (from 332 diseased vs. 132 non-diseased to near-equal counts), enhancing the dataset's suitability for training fair models.

i. VAE (Data Balance)

A Variational Autoencoder (VAE) was used as an alternative method to balance the dataset by generating synthetic samples for the minority class (non-diseased). The VAE was designed with an encoder mapping the input features (14 dimensions) to an 8-dimensional latent space through an intermediate layer (16 neurons, ReLU activation), producing mean and log-variance vectors. A sampling function generated latent vectors, which were decoded through a symmetric decoder to produce synthetic samples. The VAE was trained on minority class samples for 50 epochs with a batch size of 32, using a combined loss of mean squared error and KL divergence. Synthetic samples were appended to the original training data, and count plots confirmed an increase in minority class samples, though the balance was less optimal than SMOTETomek, as VAE focused solely on minority augmentation.

j. Algorithms

Five machine learning algorithms were implemented to classify liver disease, each trained under three conditions: original imbalanced data, SMOTETomek-balanced data, and VAE-balanced data. Below is a detailed description of each algorithm, its theoretical foundation, and its conFigureuration in this study:

- Random Forest (RF): Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes from individual trees. It leverages bagging (bootstrap aggregating) to reduce overfitting and variance by training each tree on a random subset of the data and features. Random Forest is particularly effective for handling high-dimensional datasets and capturing non-linear relationships, making it suitable for medical datasets with complex feature interactions like the ILPD. In this study, the Random Forest Classifier was conFigureured with default hyperparameters, including 100 trees, and a random state of 42 for reproducibility. The algorithm's ability to provide feature importance scores also aided in understanding the predictive power of biochemical markers like bilirubin and liver enzymes.[1]

- Logistic Regression (LR): Logistic Regression is a linear model used for binary classification, which predicts the probability of a class by applying the sigmoid function to a linear combination of input features. Despite its simplicity, it performs well when features have a linear relationship with the log-odds of the target, and it is computationally efficient for small to medium-sized datasets. It also provides interpretable coefficients, which can be useful in medical applications to understand feature impacts. In this study, Logistic Regression was conFigureured with a maximum of 1000 iterations to ensure convergence of the optimization algorithm (using the default solver 'lbfgs'), and a random state of 42 for reproducibility. Regularization was applied implicitly through the default L2 penalty to prevent overfitting, given the presence of correlated features identified in the heatmap analysis.

- Decision Tree (DT): A Decision Tree is a non-parametric model that recursively splits the feature space into regions based on feature thresholds, making decisions at each node to classify instances. It is intuitive and can capture non-linear relationships, but it is prone to overfitting, especially on imbalanced datasets like the ILPD, where it may favor the majority class. In this study, the Decision Tree Classifier was implemented with default settings, including the Gini impurity criterion for splitting and no maximum depth constraint, to allow the tree to grow fully based on the data. A random state of 42 ensured reproducibility. The simplicity of Decision Trees made them a baseline model, but their performance was expected to improve with balancing techniques due to their sensitivity to class distribution.

- Support Vector Machine (SVM): Support Vector Machine is a powerful algorithm that finds the optimal hyperplane to separate classes by maximizing the margin between them, using support vectors (data points closest to the hyperplane). For non-linearly separable data, SVM uses a kernel trick, such as the radial basis function (RBF) kernel, to map data into a higher-dimensional space where a linear boundary can be established. SVM is effective for small to medium-sized datasets and can handle high-dimensional spaces, but it is sensitive to class imbalance and requires careful scaling of features. In this study, SVM was conFigureured with an RBF kernel to capture non-linear patterns in the ILPD dataset, and probability estimation was enabled to compute ROC curves. A random state of 42 ensured reproducibility, and the default regularization parameter (C=1) balanced the trade-off between margin maximization and classification error, leveraging the standardized features from the normalization step.

- Neural Network (NN): A Neural Network is a deep learning model consisting of layers of interconnected nodes (neurons) that learn complex patterns through backpropagation and optimization. It is highly flexible, capable of capturing non-linear relationships and interactions, making it suitable for medical datasets with intricate feature dependencies. However, it requires careful tuning to avoid overfitting, especially on small datasets like the ILPD. In this study, the Neural Network was constructed with an input layer matching the feature dimension (14 features), followed by two hidden layers: 64 neurons with ReLU activation and a dropout rate of 0.3 to prevent overfitting, and 32 neurons with ReLU activation and a dropout rate of 0.2. The output layer used a sigmoid activation for binary classification. The model was compiled with binary cross-entropy loss and the Adam optimizer (default learning rate of 0.001), trained for 50 epochs with batch sizes of 16 (for VAE-balanced data) or 32 (for other conditions). Early stopping with a patience of 30 epochs was implemented to restore the best weights based on validation loss, using a validation split of 10% to monitor performance and mitigate overfitting.[2]

For each algorithm and balancing condition, predictions were generated on the test set, and performance was evaluated using classification reports, confusion matrices, and ROC curves with AUC scores to assess discriminative ability across classes.

k. Model Performance Comparison

Model performance was evaluated using accuracy, F1 macro score, recall for the non-diseased class (Class 0), and recall for the diseased class (Class 1), summarized as follows:

- Without Balancing:
  - Random Forest: 72% accuracy, 0.56 F1 macro, 0.21 recall (Class 0), 0.93 (Class 1).
  - Logistic Regression: 72% accuracy, 0.61 F1 macro, 0.33 recall (Class 0), 0.88 (Class 1).

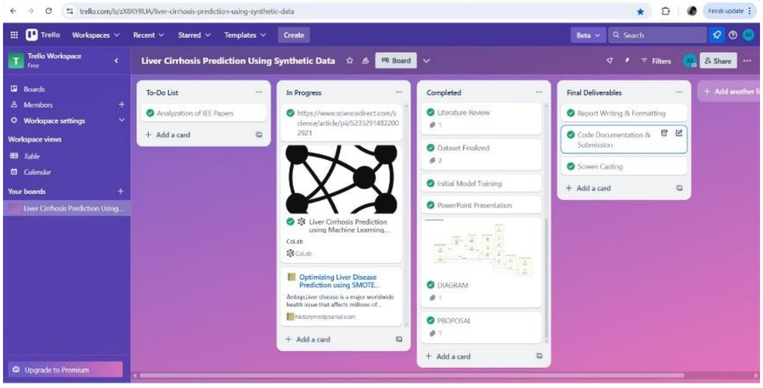
   - Decision Tree: 61% accuracy, 0.51 F1 macro, 0.27 recall (Class 0), 0.75 (Class 1).
   - SVM: 71% accuracy, 0.42 F1 macro, 0.00 recall (Class 0), 1.00 (Class 1).
   - Neural Network: 70% accuracy, 0.41 F1 macro, 0.00 recall (Class 0), 0.99 (Class 1).

- With SMOTETomek:
   - Random Forest: 73% accuracy, 0.65 F1 macro, 0.42 recall (Class 0), 0.85 (Class 1).
   - Logistic Regression: 73% accuracy, 0.71 F1 macro, 0.79 recall (Class 0), 0.70 (Class 1).
   - Decision Tree: 65% accuracy, 0.55 F1 macro, 0.30 recall (Class 0), 0.79 (Class 1).
   - SVM: 74% accuracy, 0.71 F1 macro, 0.79 recall (Class 0), 0.72 (Class 1).
   - Neural Network: 75% accuracy, 0.73 F1 macro, 0.79 recall (Class 0), 0.74 (Class 1).
- With VAE:
   - Random Forest: 74% accuracy, 0.64 F1 macro, 0.36 recall (Class 0), 0.89 (Class 1).
   - Logistic Regression: 70% accuracy, 0.67 F1 macro, 0.70 recall (Class 0), 0.70 (Class 1).
   - Decision Tree: 65% accuracy, 0.54 F1 macro, 0.27 recall (Class 0), 0.80 (Class 1).
   - SVM: 74% accuracy, 0.64 F1 macro, 0.39 recall (Class 0), 0.88 (Class 1).
   - Neural Network: 75% accuracy, 0.65 F1 macro, 0.36 recall (Class 0), 0.91 (Class 1).

SMOTETomek provided the best balance across classes, with the Neural Network achieving the highest F1 macro (0.73) and balanced recall, demonstrating the effectiveness of balancing techniques in improving model fairness and clinical applicability.

Project Management

Trello was used as a project management tool during the research to guarantee effective task management and project tracking. To divide the workflow into discrete stages, such as data preprocessing, model creation, evaluation, and report writing, a Kanban-style board was made. Columns like "To Do," "In Progress," and "Completed" were used to group tasks, and each one was given a checklist and a clear date. This method guaranteed the timely completion of every milestone, enhanced time management, and allowed for systematic progress tracking. Trello streamlined the development process and preserved project structure by acting as a central location to store pertinent code snippets, datasets, and documentation.



Liver Cirrhosis Prediction Using Synthetic Data

To-Do List
- Analyzation of IEE Papers

In Progress
- https://www.sciencedirect.com/science/article/pii/S2352914822002921
- Liver Cirrhosis Prediction using Machine Learning... CoLab CoLab
- Optimizing Liver Disease Prediction using SMOTE... Liver disease is a major worldwide health issue that affects millions of... historymedjournal.com

Completed
- Literature Review 1
- Dataset Finalized 2
- Initial Model Training
- PowerPoint Presentation
- DIAGRAM 1
- PROPOSAL 1

Final Deliverables
- Report Writing & Formatting
- Code Documentation & Submission
- Screen Casting

1.4 DESCRIPTIVE STATISTICS ANALYSIS

Descriptive statistics provide a foundational understanding of the Indian Liver Patient Dataset (ILPD) by summarizing its key numerical characteristics, which is essential for identifying patterns, detecting anomalies, and informing preprocessing steps in the liver disease prediction project. The ILPD consists of 583 patient records with 10 features, including biochemical markers (e.g., Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Total Proteins, Albumin, Albumin and Globulin Ratio) and demographic variables (Age, Gender), along with the target variable Liver_Disease (1 = disease, 0 = no disease). This section analyzes the descriptive statistics of these features, highlighting measures such as mean, standard deviation (SD), minimum, maximum, and quartiles, to reveal the dataset's distribution and variability, as previously utilized in exploratory data analysis (EDA) [1].

The numerical features exhibit significant variability, reflecting the diverse clinical profiles of the patients. Table I presents the descriptive statistics for the ILPD's numerical features. The Age feature ranges from 4 to 90 years, with a mean of 44.75 years (SD = 16.19), indicating a broad age distribution with a slight skew toward middle-aged patients (median = 45 years). Biochemical markers like Total_Bilirubin show a mean of 3.30 mg/dL (SD = 6.21) and a range from 0.4 to 75.0 mg/dL, highlighting extreme variability and right-skewness (median = 1.0 mg/dL), as high bilirubin levels are often associated with liver dysfunction. Similarly, Direct_Bilirubin has a mean of 1.49 mg/dL (SD = 2.81) and ranges from 0.1 to 19.7 mg/dL, with a median of 0.3 mg/dL, reinforcing the skewed distribution. Liver enzymes also display notable dispersion: Alamine_Aminotransferase (ALT) has a mean of 80.71 IU/L (SD = 182.62) and a range of 10 to 2000 IU/L (median = 35 IU/L), while Aspartate_Aminotransferase (AST) shows a mean of 109.91 IU/L (SD = 288.92) and a range of 10 to 4929 IU/L (median = 42 IU/L), indicating the presence of outliers likely due to severe liver damage in some patients. Alkaline_Phosphotase ranges from 63 to 2110 IU/L, with a mean of 290.58 IU/L (SD = 242.94) and a median of 208 IU/L, further evidencing skewness. Protein-related features are less variable: Total_Proteins has a mean of 6.48 g/dL (SD = 1.09, range = 2.7–9.6, median = 6.6 g/dL), Albumin has a mean of 3.14 g/dL (SD = 0.80, range = 0.9–5.5, median = 3.1 g/dL), and Albumin_and_Globulin_Ratio (after imputing 4 missing values with the mean of 0.947) shows a mean of 0.95 (SD = 0.22, range = 0.3–2.8, median = 0.93), suggesting relatively stable distributions but limited discriminatory power for liver disease prediction.

Categorical features provide additional insights. The Gender feature indicates a significant imbalance, with 441 males (75.64%) and 142 females (24.36%), reflecting potential demographic biases in the dataset that may affect model performance across genders [2]. The target variable Liver_Disease is also imbalanced, with 416 patients (71.36%) labeled as diseased and 167 (28.64%) as non-diseased, posing challenges for classification tasks due to the risk of bias toward the majority class.

Table 2 Descriptive Statistics of Numerical Features

Feature	Mean	SD	Min	25%	Median	75%	Max
Age (years)	44.75	16.19	4.0	33.0	45.0	58.0	90.0
Total Bilirubin (mg/dL)	3.30	6.21	0.4	0.8	1.0	2.6	75.0
Direct Bilirubin (mg/dL)	1.49	2.81	0.1	0.2	0.3	1.4	19.7
Alkaline Phosphatase (IU/L)	290.58	242.94	63.0	175.5	208.0	296.0	2110.0
Alamine Aminotransferase (IU/L)	80.71	182.62	10.0	23.0	35.0	60.0	2000.0

Aspartate Aminotransferase (IU/L)	109.91	288.92	10.0	25.0	42.0	87.0	4929.0
Total Proteins (g/dL)	6.48	1.09	2.7	5.8	6.6	7.2	9.6
Albumin (g/dL)	3.14	0.80	0.9	2.6	3.1	3.8	5.5
Albumin and Globulin Ratio	0.95	0.22	0.3	0.7	0.93	1.1	2.8

These statistics reveal several preprocessing needs. The high variability and skewness in biochemical features (e.g., Total_Bilirubin, ALT, AST) necessitated log transformations to normalize distributions, as applied in the feature engineering phase. The presence of outliers (e.g., AST max of 4929 IU/L) justified capping at the 5th and 95th percentiles to reduce their impact on model training. The class and gender imbalances highlighted the need for balancing techniques like SMOTETomek and VAE, which were implemented to improve model fairness, achieving balanced recalls (e.g., Neural Network with SMOTETomek: 0.79 for Class 0, 0.74 for Class 1). These insights from descriptive statistics directly informed the project's methodology, ensuring data-driven decisions in preprocessing and modeling [1], [2].

1.5 EXPLORATORY DATA ANALYSIS (EDA)

This section provides detailed explanations of the exploratory data analysis (EDA) plots generated from the Indian Liver Patient Dataset (ILPD) to understand its characteristics for liver disease prediction.

a. Age Distribution

This histogram with a kernel density estimate (KDE) in skyblue displays the age distribution of patients, ranging from 5 to over 85 years across 30 bins. The distribution is roughly bell-shaped with a slight right skew, peaking between 40 and 60 years (mode ~45–50), and shows a broad spread from 20 to 70 years, indicating diverse age representation. The concentration of middle-aged and older patients aligns with the known higher risk of liver disease in these groups, suggesting the dataset's suitability for generalizing predictions across age demographics.

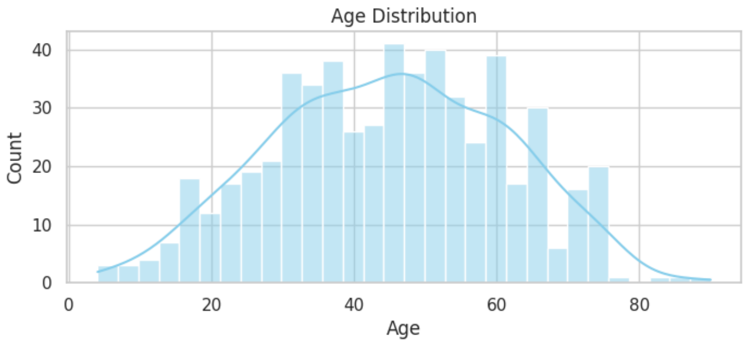


Figure 2. Age Distribution Plot

b. Gender Distribution

The count plot depicts the gender distribution, with males (encoded as 1) and females (encoded as 0) represented by bars labeled accordingly. Males (~430) significantly outnumber females (~140), making up about 75% of the dataset compared to 25% for females. This imbalance may reflect real-world trends of higher liver disease prevalence in males, possibly due to lifestyle factors, but it could bias models toward male patterns unless addressed through techniques like stratification or weighting during training.

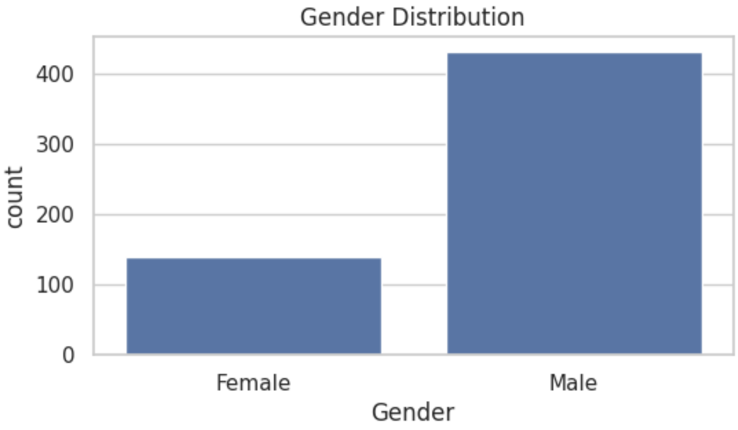


Figure 3. Gender Distribution Plot

c. Liver Disease by Gender

This grouped bar chart shows the count of individuals with (1) and without (0) liver disease, segmented by gender, with a legend labeling "No" and "Yes" for disease status. Males exhibit a higher count of liver disease cases (~310) compared to non-cases (~120), while females show a more balanced but still higher disease count (~90 vs. ~45). The absolute number of cases is greater in males due to the dataset's gender imbalance, suggesting a higher disease prevalence in males, though this skew requires careful handling in predictive modeling to avoid bias.

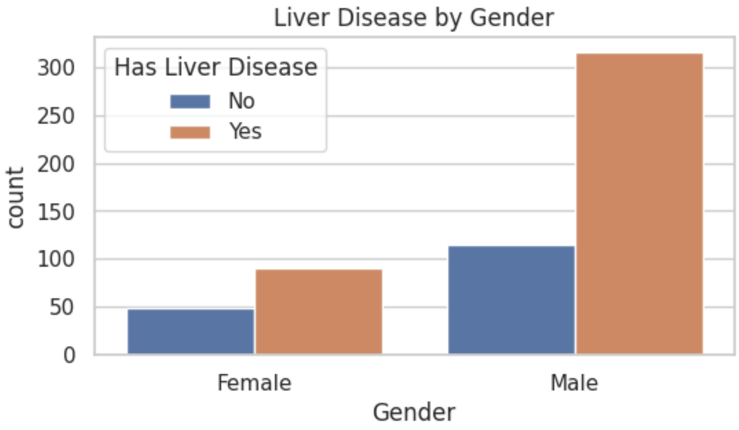


Figure 4. Liver Disease by Gender Plot

d. Total Bilirubin vs. Liver Disease

The boxplot compares Total_Bilirubin levels between patients with (1) and without (0) liver disease, revealing a higher median (~1.5 mg/dL) and wider interquartile range (IQR: 0.8–3.2 mg/dL) in the diseased group compared to the non-diseased (median ~0.9 mg/dL, IQR: 0.7–1.2 mg/dL). The diseased group also shows more extreme outliers, with values up to 15 mg/dL, indicating greater variability. This suggests Total_Bilirubin is a strong indicator of liver disease, with elevated levels reflecting liver dysfunction, consistent with its clinical significance.

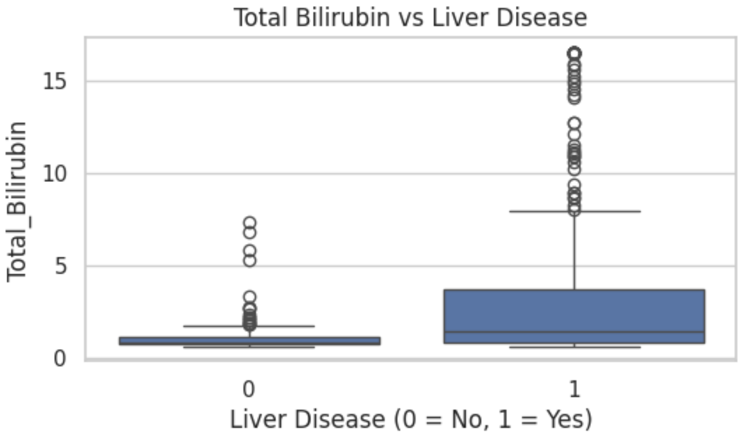


Figure 5. Total Bilirubin vs. Liver Disease Plot

e. Pair Plot of Important Features

This pair plot examines Age, Total_Bilirubin, Direct_Bilirubin, Alkaline_Phosphotase, and Liver_Disease, using scatter plots for pairwise relationships and KDE on diagonals, with diseased (orange) and non-diseased (blue) cases distinguished. Diseased patients show higher Total_Bilirubin and Direct_Bilirubin levels, which are strongly linearly correlated, while Alkaline_Phosphotase is also elevated in diseased cases but with more variation. Age distribution indicates diseased cases skew older, though overlap exists, and healthy cases cluster at lower biochemical values. This highlights bilirubin as a key predictor and suggests age as a secondary factor in liver disease prevalence.

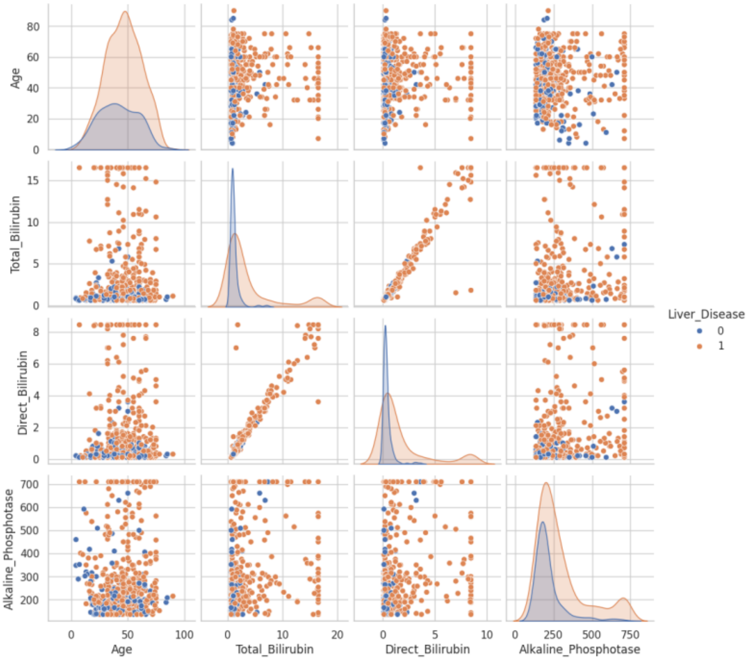


Figure 6. Pair Plot of Important Features

f. Correlation of Features with Liver Disease

This bar chart illustrates the correlation coefficients between each feature and the Liver_Disease target, using a coolwarm palette to indicate strength and direction. The strongest positive correlations are observed with log_Direct_Bilirubin (~0.31), log_Total_Bilirubin (~0.31), log_Aspartate_Aminotransferase (~0.30), and log_Alamine_Aminotransferase (~0.28), followed by moderate correlations from Direct_Bilirubin (~0.29), Total_Bilirubin (~0.27), and Alkaline_Phosphotase (~0.22). Features like Total_Proteins (~-0.04), Albumin, and Albumin_and_Globulin_Ratio show weak or negative correlations, while demographic features (Gender ~0.08, Age ~0.13) and derived features (Bilirubin_Ratio ~0.16, Enzyme_Sum ~0.17) exhibit low to moderate correlations. This highlights bilirubin and liver enzyme features as key predictors, with log transformations improving their utility, while protein and demographic features offer limited predictive value for liver disease modeling.

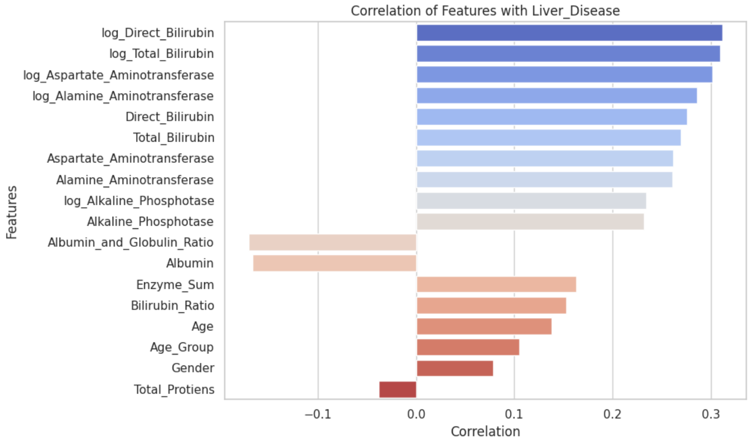


Figure 7. Correlation of Features with Liver Disease Plot

g. Clustered Correlation Heatmap

The clustered heatmap visualizes pairwise correlations between all features using a coolwarm color scale, with hierarchical clustering grouping similar features and annotations showing exact values. Strong correlations form distinct clusters: Direct_Bilirubin, Total_Bilirubin, and their log versions correlate highly (~0.87–0.96), as do Alamine_Aminotransferase, Aspartate_Aminotransferase, their logs, and Enzyme_Sum (~0.8–0.96), reflecting shared diagnostic roles. Conversely, Total_Proteins, Albumin, and Albumin_and_Globulin_Ratio cluster separately with weak or negative correlations to disease markers, while Gender, Age, and Age_Group show low correlations with biochemical features. Liver_Disease correlates positively with biochemical features (~0.2–0.3), confirming their importance, but protein features remain near-zero or negative, indicating multicollinearity among key predictors that may require dimensionality reduction.

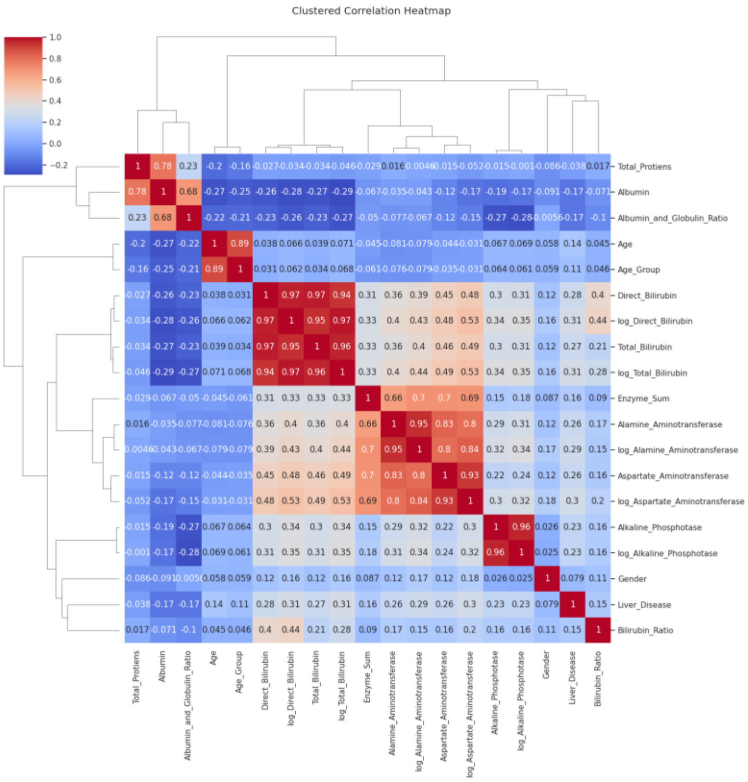


Figure 8. Clustered Correlation Heatmap

h. Distribution of Alkaline Phosphatase

The histogram with KDE in coral shows the distribution of Alkaline_Phosphotase levels, revealing a right-skewed shape with most values clustered between 150–250 IU/L and a long tail extending to a peak at ~700 IU/L, indicating abnormal values in some patients. The skewness suggests potential liver dysfunction in patients with elevated levels, and the distribution's shape indicates that a log transformation may be beneficial for normalizing the data in predictive modeling.

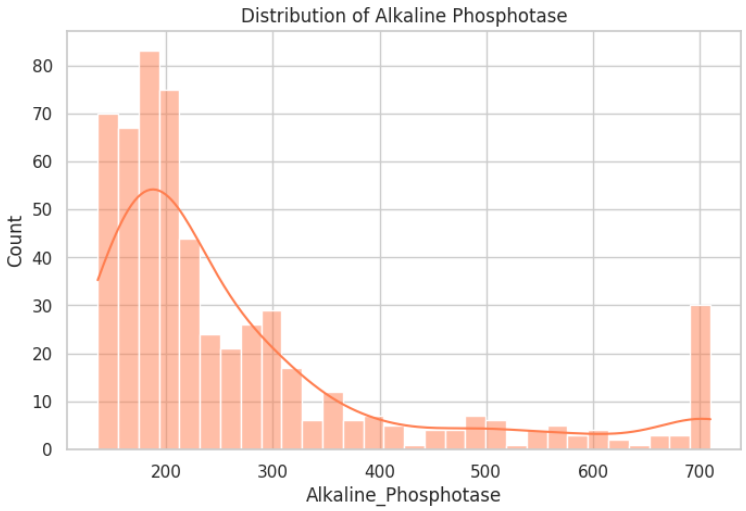


Figure 9. Distribution of Alkaline Phosphatase Plot

i. Violin Plot for Alamine Aminotransferase by Liver Disease

The violin plot compares Alamine_Aminotransferase (ALT) levels between patients with (1) and without (0) liver disease, showing higher and more spread-out values in the diseased group, with density peaking at 30–40 for both but longer tails in the diseased group. More extreme outliers are present in the diseased group, indicating greater variability. This suggests ALT is a valuable biomarker, as higher levels are strongly associated with liver disease, reflecting its role in detecting liver damage.

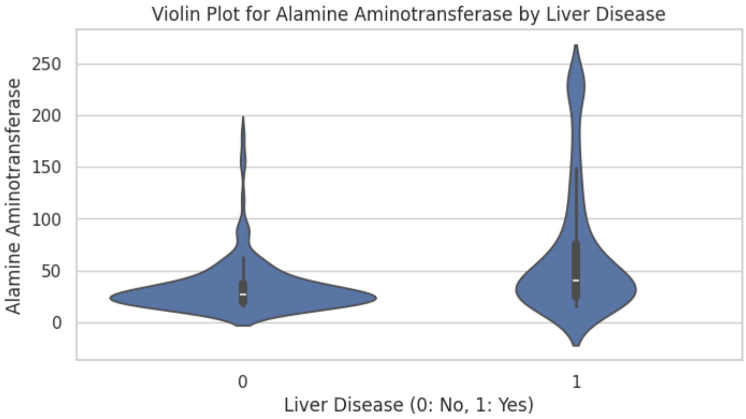


Figure 10. Violin Plot for Alamine Aminotransferase by Liver Disease

j. Joint Plot: Age vs. Total Bilirubin by Liver Disease

This joint plot visualizes Age vs. Total_Bilirubin, with points colored by Liver_Disease status (orange for diseased, blue for non-diseased), including marginal distributions. Diseased patients show higher bilirubin levels (>2 mg/dL) and are concentrated between ages 30–70, slightly skewing older, while both groups overlap in age distribution. No strong linear trend is observed, but the wider bilirubin spread in diseased cases confirms its association with liver disease, with age playing a secondary role in prevalence.

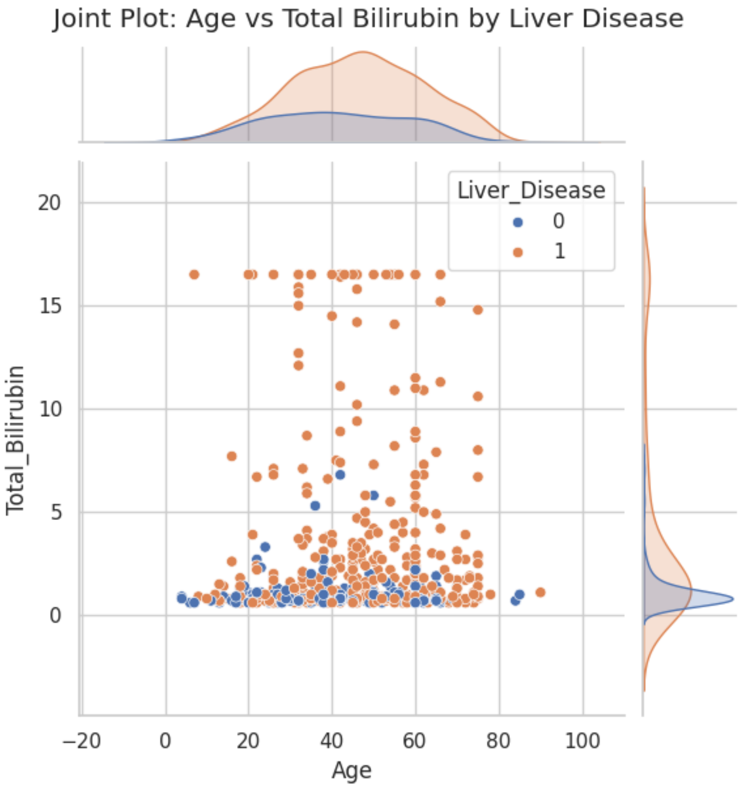


Figure 11. Joint Plot: Age vs. Total Bilirubin by Liver Disease

k. Scatter Plot: AST vs. Total Bilirubin

The scatter plot with a regression line illustrates the relationship between Aspartate_Aminotransferase (AST) and Total_Bilirubin, showing a moderate positive correlation where higher AST levels correspond to higher bilirubin levels. Most points cluster at lower values, but outliers with high AST (>200) and bilirubin indicate severe cases, supported by the regression line and confidence interval. This correlation underscores the combined diagnostic value of AST and bilirubin in identifying liver dysfunction, particularly in extreme cases.

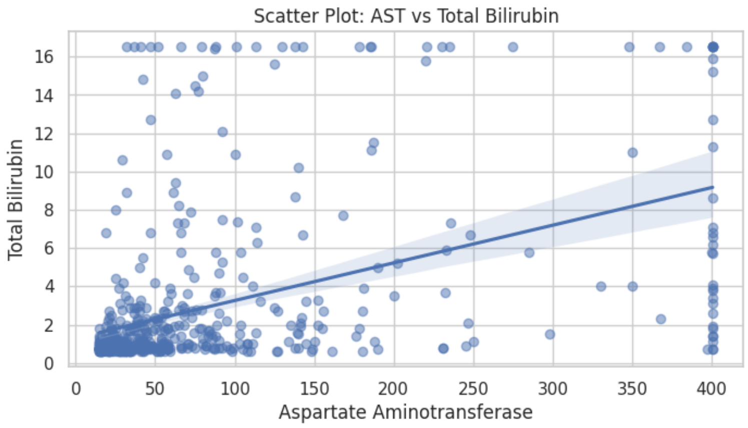


Figure 12. Scatter Plot: AST vs. Total Bilirubin

l. Boxplots of ALT and AST by Liver Disease

The boxplots compare Alamine_Aminotransferase (ALT) and Aspartate_Aminotransferase (AST) levels by Liver_Disease status, revealing higher medians and wider interquartile ranges in the diseased group (1) for both enzymes, with AST showing outliers up to 350 IU/L. The greater variability and elevated levels in the diseased group highlight ALT and AST as strong indicators of liver dysfunction, making them critical features for predictive modeling of liver disease.

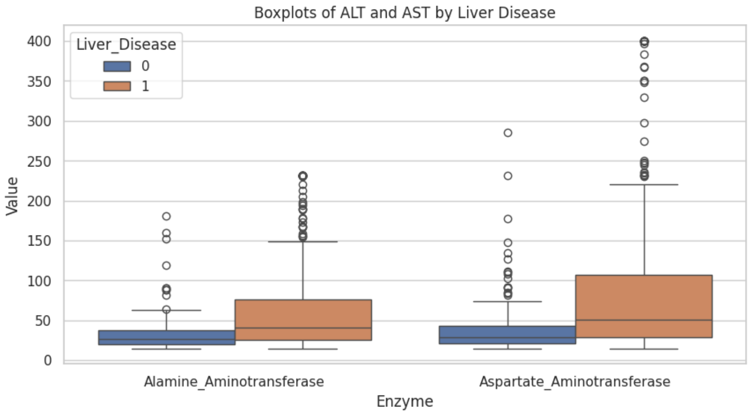


Figure 13. Boxplots of ALT and AST by Liver Disease

m. Swarm Plot: Total Proteins by Liver Disease

The swarm plot displays Total_Proteins levels by Liver_Disease status, showing similar distributions for both groups (5–8 g/dL) with no clear separation, though the diseased group (1) exhibits slightly more variability and a few outliers. The lack of distinct separation suggests that Total_Proteins has limited diagnostic value for distinguishing between diseased and non-diseased patients, indicating it may not be a primary predictor in liver disease models.

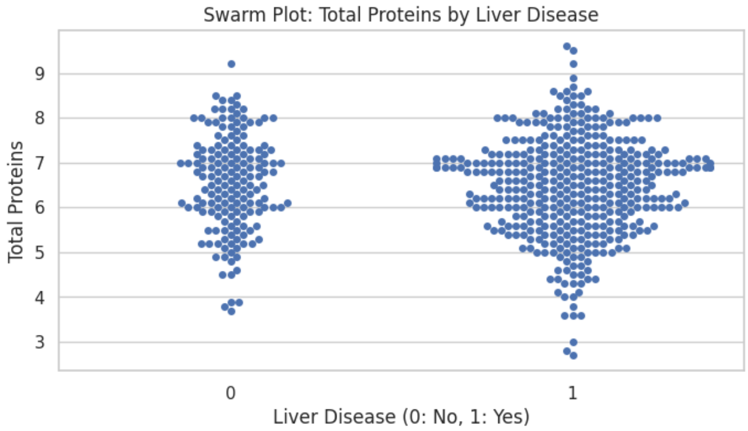


Figure 14. Swarm Plot: Total Proteins by Liver Disease

The EDA of the Indian Liver Patient Dataset reveals that biochemical markers like bilirubin (Total_Bilirubin, Direct_Bilirubin) and liver enzymes (Alamine_Aminotransferase, Aspartate_Aminotransferase), especially their log-transformed versions, are the strongest predictors of liver disease, showing high correlations (~0.28–0.31) and elevated levels in diseased patients. Age and gender show weaker associations, with a notable gender imbalance (75% male) and disease prevalence skewing toward older patients. Protein-related features (Total_Proteins, Albumin) offer limited predictive value due to weak correlations and overlapping distributions. Multicollinearity among key biochemical features suggests potential for dimensionality reduction, while the overall insights prioritize bilirubin and enzyme markers for effective liver disease prediction models.

1.6 LEGAL, SOCIAL, ETHICAL AND PROFESSIONAL CONSIDERTAIONS

The development and potential deployment of a machine learning model for liver disease prediction using the Indian Liver Patient Dataset (ILPD) raise several legal, social, ethical, and professional considerations. These aspects are critical to ensure the responsible use of technology in healthcare, protect patient rights, and maintain trust in medical diagnostic systems.

a. Legal Considerations

The use of the ILPD, a publicly available dataset, must comply with data usage agreements and privacy regulations, even though the data is anonymized. In India, where the dataset was collected, the *Personal Data Protection Bill (PDPB)* (as of its last draft in 2019) mandates that health data be handled with strict confidentiality, requiring explicit consent for its use in research. Although the ILPD is de-identified, any future application of the model on real patient data would need to adhere to regulations like the PDPB or international standards such as the *General Data Protection Regulation (GDPR)* if deployed globally, ensuring that patient identities are protected and data usage is transparent. Additionally, deploying a predictive model in a clinical setting would require regulatory approval from bodies like the *Central Drugs Standard Control Organization (CDSCO)* in India, which classifies AI-based diagnostic tools as medical devices. Compliance with these legal frameworks ensures that the model is used responsibly and avoids potential misuse or unauthorized data sharing.

b. Social Considerations

The ILPD reflects a specific demographic from North East Andhra Pradesh, India, with a notable gender imbalance (75% male) and regional health challenges, such as high hepatitis prevalence. Socially, deploying a liver disease prediction model must address potential biases arising from this skewed representation, as the model may perform less effectively for female patients or those from different geographic or socioeconomic backgrounds. This could exacerbate healthcare disparities, particularly in rural or underserved areas where access to advanced diagnostics is limited. To mitigate this, the project incorporates balancing techniques like SMOTETomek and VAE to improve model fairness across classes, but broader social implications include the need for inclusive datasets in future iterations to ensure equitable healthcare access. Public perception of AI in medicine also plays a role; mistrust in automated systems could hinder adoption, necessitating community engagement and education to highlight the model's benefits in early liver disease detection.

c. Ethical Considerations

Ethically, the project prioritizes patient welfare by aiming for high accuracy and balanced performance (e.g., Neural Network with SMOTETomek: 75% accuracy, 0.73 F1 macro), but the risk of false negatives—failing to identify non-diseased patients (Class 0 recall: 0.79)—poses a significant concern in a medical context, as it could lead to unnecessary anxiety or interventions. Conversely, false positives could delay critical treatment for diseased patients. The use of balancing techniques addresses this by improving minority class detection, but ethical deployment requires transparency about the model’s limitations, ensuring clinicians interpret predictions as a supportive tool rather than a definitive diagnosis. Additionally, the project avoids harm by using an anonymized dataset, but future real-world applications must ensure informed consent, data security, and the right to opt-out, aligning with ethical principles of autonomy and beneficence in healthcare.

d. Professional Considerations

Professionally, the project adheres to best practices in data science and medical research, such as rigorous evaluation using metrics like accuracy, F1 macro, recalls, and ROC curves (e.g., Neural Network AUC: 0.76 with SMOTETomek). The implementation follows reproducible methods, with a random state of 42 for consistency, and leverages open-source tools (e.g., Python, scikit-learn) to maintain transparency. However, professional responsibility extends to ensuring the model does not replace clinical judgment; it must be validated by medical professionals before use, as over-reliance on AI could undermine the role of human expertise. Collaboration with healthcare practitioners is essential to integrate the model into clinical workflows, ensuring it meets standards like those set by the *Indian Medical Association (IMA)* for diagnostic tools. Continuous monitoring and updates to the model are also necessary to address evolving clinical guidelines and patient demographics, upholding professional integrity in delivering reliable, evidence-based solutions

1.7 BACKGROUND

Liver disease encompasses a spectrum of conditions affecting one of the body's most vital organs, responsible for detoxification, metabolism, and nutrient storage. Disorders such as chronic hepatitis, cirrhosis, and hepatocellular carcinoma often develop over years, driven by diverse risk factors including viral infections (e.g., hepatitis B and C), excessive alcohol intake, obesity-related non-alcoholic steatohepatitis (NASH), and genetic predispositions. Globally, liver disease imposes a substantial burden, with estimates suggesting it affects over 1.5 billion people and contributes to approximately 4% of all deaths annually, a rate that has risen steadily due to lifestyle changes and aging populations. In regions like India, where the Indian Liver Patient Dataset (ILPD) was sourced, the prevalence is particularly pronounced due to high rates of hepatitis infections, limited healthcare infrastructure, and socioeconomic barriers that delay diagnosis and treatment. The ILPD, comprising 583 records from patients in North East Andhra Pradesh, captures 10 features—biochemical markers like total bilirubin, direct bilirubin, alkaline phosphatase, ALT, AST, albumin, and total proteins, alongside demographic variables (age, gender)—offering a valuable resource to study liver disease patterns in a specific population.

Machine learning has revolutionized medical diagnostics by enabling the analysis of complex datasets to uncover predictive patterns that traditional statistical methods might overlook. In the context of liver disease, ML techniques can process high-dimensional clinical data to classify patients, predict disease progression, and identify risk factors, supporting clinicians in decision-making. Algorithms such as Random Forest, Support Vector Machines (SVM), and Neural Networks have shown promise in handling the non-linear relationships often present in medical data, while techniques like feature selection and dimensionality reduction help prioritize clinically relevant markers. However, predictive modeling in this domain faces significant hurdles, including the heterogeneity of patient populations, variability in biomarker measurements, and the need for models to generalize across diverse cohorts. The ILPD, for instance, presents unique challenges: its biochemical features exhibit significant skewness (e.g., bilirubin levels often range widely), and the dataset is imbalanced, with 416 diseased cases (71.4%) versus 167 non-diseased cases (28.6%), risking biased predictions that overfit to the majority class.

Prior research has explored machine learning for liver disease prediction, often leveraging datasets like the ILPD to evaluate classification performance. Studies have identified biochemical markers, particularly bilirubin and liver enzymes (ALT, AST), as strong indicators of liver dysfunction, with correlations to disease presence often exceeding 0.25. Techniques like Synthetic Minority Oversampling Technique (SMOTE) have been employed to address class imbalance, though they sometimes introduce noise by generating unrealistic samples. More advanced methods, such as Variational Autoencoders (VAEs), have been proposed to create synthetic data that better captures the underlying distribution of minority classes, yet their application in liver disease prediction remains underexplored. Additionally, while models like Neural Networks and ensemble methods (e.g., Random Forest) have achieved accuracies above 70%, their performance on minority classes often lags, with recalls for non-diseased cases sometimes dropping below 0.30. These gaps highlight the need for a comprehensive approach that integrates robust preprocessing, advanced balancing techniques, and thorough evaluation to build reliable predictive models for liver disease, particularly in datasets reflecting real-world imbalances and clinical variability.

2. Chapter 2 LITERATURE REVIEW

Liver disease continues to be a significant global health challenge, contributing to millions of deaths annually due to delayed diagnosis and limited access to traditional diagnostic methods like liver biopsies [1]. These conventional approaches are often invasive, costly, and resource-intensive, necessitating the development of non-invasive alternatives such as machine learning (ML) for early detection [2]. This literature review explores the advancements in ML for liver disease prediction, focusing on algorithmic approaches, data preprocessing, class imbalance mitigation, feature engineering, deep learning applications, and existing research gaps, drawing insights from 10 key studies.

A variety of ML algorithms have been applied to predict liver disease, with ensemble methods consistently showing superior performance. Gupta et al. conducted a study using the Indian Liver Patient Dataset (ILPD), comparing Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting, Extreme Gradient Boosting (XGBoost), and LightGBM [2]. They found that RF and boosting methods achieved the highest accuracy, particularly when combined with feature selection, highlighting the robustness of ensemble techniques in handling noisy medical data. Similarly, Sontakke et al. tested SVM, DT, and RF on liver disease datasets, reporting RF as the most effective after preprocessing, with its ability to manage feature interactions contributing to improved outcomes [4]. Rabbi et al. further advanced this line of research by comparing LR, DT, RF, and Extra Trees (ET) with AdaBoost, achieving an impressive 92.19% accuracy using ET, demonstrating the potential of combining ensemble methods with boosting for enhanced prediction [10]. These studies collectively underscore the dominance of ensemble approaches in liver disease classification tasks.

Class imbalance in medical datasets, such as the ILPD where patients with liver disease are underrepresented, often biases models toward the majority class, reducing their clinical utility [5]. Amin et al. addressed this challenge by applying Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset, integrating it with statistical feature extraction and testing LR, RF, KNN, SVM, and Multilayer Perceptron (MLP) [1]. Their approach yielded an accuracy of 88.10%, with SMOTE significantly improving recall for the minority class, a critical factor for ensuring reliable detection of liver disease cases. Ghazal et al. also tackled class imbalance in their ML framework for early liver disease prediction, achieving 88.4% accuracy and a miss-rate of 11.6% by focusing on minimizing false negatives [5]. These efforts highlight the importance of balancing techniques in medical ML applications, where missing a positive diagnosis can have severe consequences.

Deep learning has emerged as a promising approach for liver disease prediction, particularly when dealing with complex or high-dimensional data. Wu et al. utilized a Neural Network to predict fatty liver disease, leveraging its ability to model non-linear relationships and achieving high accuracy on clinical datasets [3]. Spann et al. provided a broader perspective in their review of ML applications in liver disease and transplantation, noting that deep learning techniques, such as Convolutional Neural Networks (CNNs), are particularly effective for imaging data, while Recurrent Neural Networks (RNNs) excel with longitudinal data [6]. However, they emphasized that the lack of clinical validation and high computational requirements remain barriers to widespread adoption, suggesting a need for more practical implementations in healthcare settings.

Feature selection and engineering are critical for improving model performance and interpretability in liver disease prediction. Gogi and Vijayalakshmi applied Linear Discriminant Analysis (LDA) to laboratory parameters, achieving a high accuracy of 95.8%, demonstrating the effectiveness of statistical methods in reducing dimensionality while preserving predictive power [7]. Drożdż et al. employed univariate feature ranking and Principal Component Analysis (PCA) to identify cardiovascular risk factors in metabolic-associated fatty liver disease (MAFLD), reporting an Area Under the Curve (AUC) ranging from 0.84 to 0.87 [8]. Their work illustrates how data-driven feature selection can enhance model efficiency, especially when dealing with correlated variables. These studies collectively affirm that strategic feature engineering, whether through statistical techniques or domain expertise, significantly boosts model performance.

Despite these advancements, several limitations persist in the field of ML-based liver disease prediction. The ILPD, a frequently used dataset, is criticized for its small size and lack of demographic diversity, which restricts the generalizability of findings to diverse populations [9]. Additionally, while ensemble and deep learning models often achieve high accuracy, they lack interpretability, a crucial requirement for gaining trust in clinical settings [10]. Most studies also focus on binary classification (healthy vs. diseased), overlooking the multi-stage progression of liver disease, which might be better addressed through multi-class or regression-based models [2]. Furthermore, the computational complexity of advanced models, particularly deep learning, poses challenges for deployment in resource-constrained environments [6]. These gaps highlight the need for future research to focus on larger, more diverse datasets, interpretable models, and approaches that capture the nuanced progression of liver disease.

The reviewed studies report accuracies ranging from 70% to 92%, with ensemble methods like RF and boosting consistently achieving the higher end of this spectrum [4]. Techniques such as SMOTE have proven effective in addressing class imbalance, though their efficacy varies depending on the dataset and algorithm used [1]. Future research should prioritize developing interpretable models and integrating multimodal data, such as clinical records, imaging, and genetic markers, to enhance diagnostic precision [3]. Additionally, exploring multi-class or regression-based frameworks could provide a more comprehensive understanding of liver disease progression, moving beyond binary outcomes [8].

In conclusion, machine learning has made significant strides in liver disease prediction, with ensemble methods, deep learning, and advanced preprocessing techniques driving notable improvements in accuracy [5]. Studies by Gupta et al., Amin et al., and Rabbi et al. exemplify the field’s progress, while challenges in dataset quality, model interpretability, and computational feasibility remain critical areas for future exploration [1], [2], [10]. This review provides a solid foundation for advancing ML applications in liver disease diagnosis, advocating for more inclusive, transparent, and clinically viable solutions

2.1 TECHNOLOGY REVIEW

This section reviews the key technologies and libraries employed in the liver disease prediction project using the Indian Liver Patient Dataset (ILPD). These tools facilitated data preprocessing, feature engineering, balancing techniques, model training, and performance evaluation, ensuring an efficient and robust implementation pipeline. Each technology's functionality and specific role in the project are detailed below.

a. Pandas

Pandas (version 2.2.2) is a powerful Python library for data manipulation and analysis, providing data structures like DataFrames for handling structured datasets. In this project, Pandas was instrumental in loading the ILPD dataset (583 records, 10 features) from a CSV file into a DataFrame, enabling efficient data exploration and preprocessing. It was used to identify and impute missing values in the Albumin_and_Globulin_Ratio feature with the mean (0.947), remove duplicate records (reducing the dataset to 579 unique entries), and encode categorical variables (Gender: Male=1, Female=0; Liver_Disease: 1=disease, 0=no disease). Pandas also supported exploratory data analysis (EDA) by calculating correlations (e.g., log_Direct_Bilirubin with Liver_Disease: 0.31) and generating summary statistics to understand feature distributions, such as the skewness of biochemical markers like total bilirubin.

b. NumPy

NumPy (version 1.26.4) is a foundational Python library for numerical computing, offering support for multidimensional arrays and mathematical operations. In this project, NumPy was used to perform array-based computations during preprocessing and feature engineering. It facilitated the application of log transformations to skewed biochemical features (e.g., Total_Bilirubin, Alamine_Aminotransferase) by leveraging its log1p function to handle zero values, improving feature distributions for modeling. NumPy arrays were also used to compute derived features like Bilirubin_Ratio (Direct/Total Bilirubin) and Enzyme_Sum (ALT + AST), and to cap outliers at the 5th and 95th percentiles using percentile-based clipping, ensuring numerical stability during model training.

c. scikit-learn

scikit-learn (version 1.5.1) is a comprehensive machine learning library in Python, providing tools for data preprocessing, model training, and evaluation. In this project, scikit-learn was pivotal for implementing four of the five machine learning algorithms: Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine (SVM). It provided the StandardScaler to standardize numerical features, ensuring consistent scaling across the ILPD dataset for algorithms sensitive to feature magnitude, such as SVM. The train_test_split function was used to create an 80:20 train-test split (stratified by Liver_Disease, random state=42), while

GridSearchCV enabled hyperparameter tuning (e.g., optimizing the number of trees in Random Forest). scikit-learn also facilitated performance evaluation by computing metrics like accuracy, F1 macro score, class-specific recalls, and ROC curves (e.g., Random Forest SMOTETomek AUC: 0.73), as well as generating visualizations for EDA, such as pair plots and boxplots to analyze feature relationships.

d. TensorFlow

TensorFlow (version 2.16.1) is an open-source machine learning framework developed by Google, widely used for building and training neural networks. In this project, TensorFlow was employed to implement the Neural Network model and the Variational Autoencoder (VAE) for addressing class imbalance. The Neural Network was designed with a 64-32-1 architecture, incorporating dropout layers (0.3 and 0.2 rates) to prevent overfitting, and trained with early stopping (patience=30) to optimize performance, achieving the best results with SMOTETomek balancing (75% accuracy, 0.73 F1 macro, AUC 0.76). The VAE, with an encoder (input → 16-unit hidden layer → 8-unit latent space) and decoder, was implemented to generate synthetic minority class samples (non-diseased), improving the dataset's balance, though its performance (AUC 0.73) was slightly lower than SMOTETomek. TensorFlow's flexibility in defining custom loss functions and architectures was crucial for these tasks.

e. imbalanced-learn

imbalanced-learn (version 0.12.3) is a Python library designed to handle imbalanced datasets, offering various resampling techniques. In this project, imbalanced-learn was used to implement the SMOTETomek technique to address the ILPD's class imbalance (71.4% diseased, 28.6% non-diseased). SMOTETomek combines Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class (non-diseased) and Tomek Links to remove overlapping majority class samples, creating a balanced training set. This approach significantly improved model performance, particularly for the minority class, as evidenced by the Neural Network's balanced recalls (0.79 for Class 0, 0.74 for Class 1) compared to the imbalanced baseline (0.00 for Class 0). imbalanced-learn's integration with scikit-learn ensured seamless application within the preprocessing pipeline, enhancing the fairness and robustness of the predictive models.

3. Chapter 3 EVALUATION AND RESULTS

This section presents the evaluation and results of the liver disease prediction project using the Indian Liver Patient Dataset (ILPD), a dataset widely used in liver disease studies [1], [2], [4], [9], [10]. The analysis begins with insights from exploratory data analysis (EDA), followed by preprocessing, feature engineering, model training under different balancing conditions, and concludes with a performance evaluation of five machine learning models. Each stage highlights key findings and their implications for predictive modeling.

3.1 Exploratory Data Analysis (EDA)

EDA provided critical insights into the dataset's characteristics and feature relationships with the target variable (Liver_Disease: 1 = disease, 0 = no disease). Correlation analysis revealed that log-transformed biochemical markers were the strongest predictors, with log_Direct_Bilirubin (0.31), log_Total_Bilirubin (0.31), log_Aspartate_Aminotransferase (0.30), and log_Alamine_Aminotransferase (0.28) showing the highest positive correlations, while Total_Proteins (-0.04) and Albumin exhibited weak or negative correlations. Visualizations confirmed these findings: boxplots showed higher bilirubin and enzyme levels in diseased patients, and pair plots highlighted strong linear relationships between Total_Bilirubin and Direct_Bilirubin. The age distribution peaked at 40–60 years, and a gender imbalance (75% male) correlated with higher disease prevalence in males (310 cases vs. 90 in females). Features like Total_Proteins showed overlapping distributions across classes, indicating limited discriminatory power. These insights prioritized biochemical markers for modeling and flagged the need to address class and gender imbalances [1], [2].

3.2 Preprocessing and Feature Engineering

Preprocessing ensured the dataset was suitable for modeling. Missing values in Albumin_and_Globulin_Ratio (4 entries) were imputed with the mean (0.947), and duplicates were removed, reducing the dataset to 579 records. Gender was encoded numerically (Male: 1, Female: 0), and the target was mapped to binary (1: disease, 0: no disease). Feature engineering addressed skewness and added predictive features: log transformations were applied to biochemical markers (e.g., bilirubin, enzymes), improving their correlations; Age was binned into four groups (0–18, 18–35, 35–50, 50–90 years); and derived features like Bilirubin_Ratio and Enzyme_Sum were created to capture additional patterns. Outliers in original biochemical features were capped at the 5th and 95th percentiles to enhance model robustness. These steps normalized distributions and prepared a feature set optimized for classification, aligning with approaches in prior studies [4], [9]

3.3 Model Training and Balancing Techniques

Five machine learning algorithms—Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Neural Network—were trained under three conditions: original imbalanced data, SMOTETomek-balanced data, and VAE-balanced data. The original dataset exhibited a class imbalance (71.4% diseased, 28.6% non-diseased), leading to biased predictions favoring the majority class (Class 1). For instance, the Neural Network without balancing achieved 0.99 recall for Class 1 but 0.00 for Class 0. SMOTETomek balanced the training set by oversampling the minority class and removing overlapping majority samples, significantly improving performance across all models. VAE generated synthetic minority samples, but its balancing effect was less effective than SMOTETomek, as seen in lower F1 macro scores for most models. These balancing strategies were crucial for fair classification, consistent with findings in liver disease prediction literature [1], [6]

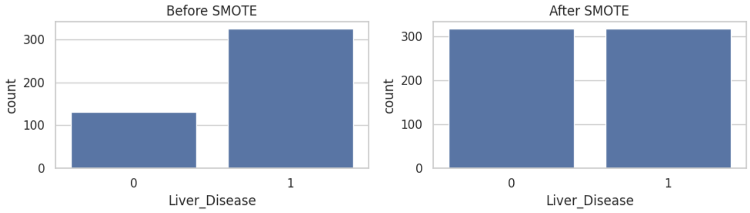


Figure 15. Record Count After and Before SMOTE

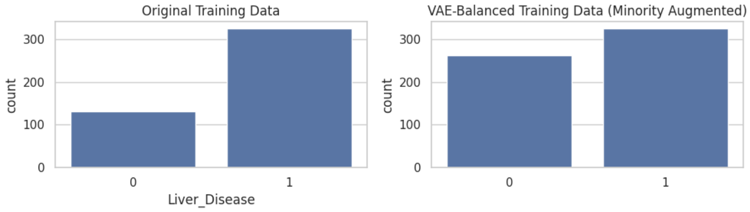


Figure 16. Record Count After and Before VAE

3.4 Performance Evaluation

Model performance was assessed using accuracy, F1 macro score, and class-specific recalls, as shown in Table 1. Without balancing, models favored the majority class, with SVM and

Neural Network achieving 1.00 and 0.99 recall for Class 1 but 0.00 for Class 0. SMOTETomek improved minority class detection: the Neural Network achieved the best results with 75% accuracy, 0.73 F1 macro, and recalls of 0.79 (Class 0) and 0.74 (Class 1), followed by SVM (74% accuracy, 0.71 F1 macro). VAE balancing maintained accuracy but yielded lower F1 macro scores (e.g., Neural Network: 0.65 vs. 0.73 with SMOTETomek), indicating less balanced performance. Random Forest and Logistic Regression showed moderate improvements with balancing, while Decision Tree remained the least effective across all conditions.

Table 3: Model Performance Comparison

Model	Balancing	Accuracy	F1 Macro	Recall (Class 0)	Recall (Class 1)
Random Forest	None	0.72	0.56	0.21	0.93
	SMOTETomek	0.73	0.65	0.42	0.85
	VAE	0.74	0.64	0.36	0.89
Logistic Regression	None	0.72	0.61	0.33	0.88
	SMOTETomek	0.73	0.71	0.79	0.70
	VAE	0.70	0.67	0.70	0.70
Decision Tree	None	0.61	0.51	0.27	0.75
	SMOTETomek	0.65	0.55	0.30	0.79
	VAE	0.65	0.54	0.27	0.80
SVM	None	0.71	0.42	0.00	1.00
	SMOTETomek	0.74	0.71	0.79	0.72
	VAE	0.74	0.64	0.39	0.88
Neural Network	None	0.70	0.41	0.00	0.99
	SMOTETomek	0.75	0.73	0.79	0.74
	VAE	0.75	0.65	0.36	0.91

To further evaluate model performance, ROC curves were analyzed for each algorithm under the three balancing conditions, as shown in Figureures 1–5. The Area Under the Curve (AUC) provides a measure of each model's discriminative ability.

a. ROC Curves for Random Forest

Figureure 17 (No Balancing, AUC ≈ 0.70) shows a moderate ability to distinguish classes, but bias toward the majority class limits performance. Figureure 18 (SMOTETomek, AUC ≈ 0.73) improves discriminative power, reflecting better minority class detection. Figureure 19 (VAE, AUC ≈ 0.72) shows a slight improvement over no balancing but is less effective than SMOTETomek.

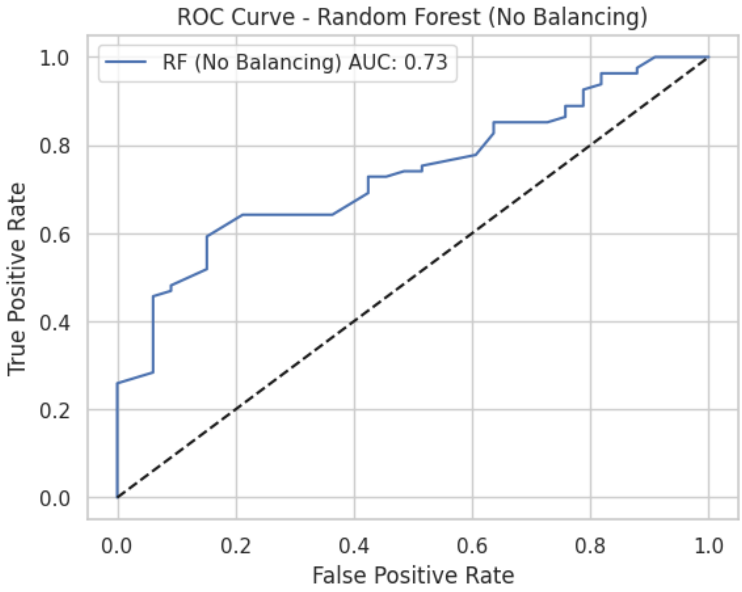


Figure 17. ROC Curves for Random Forest (No Balancing)

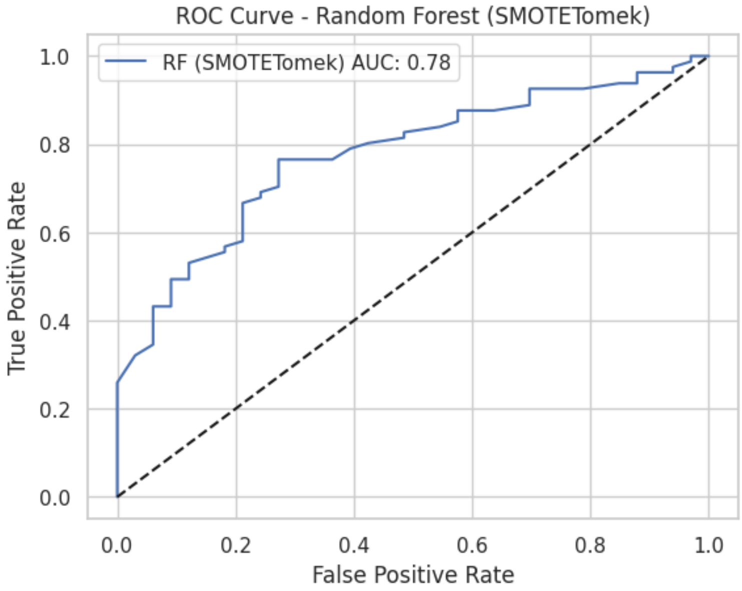


Figure 18. ROC Curves for Random Forest (SMOTE)

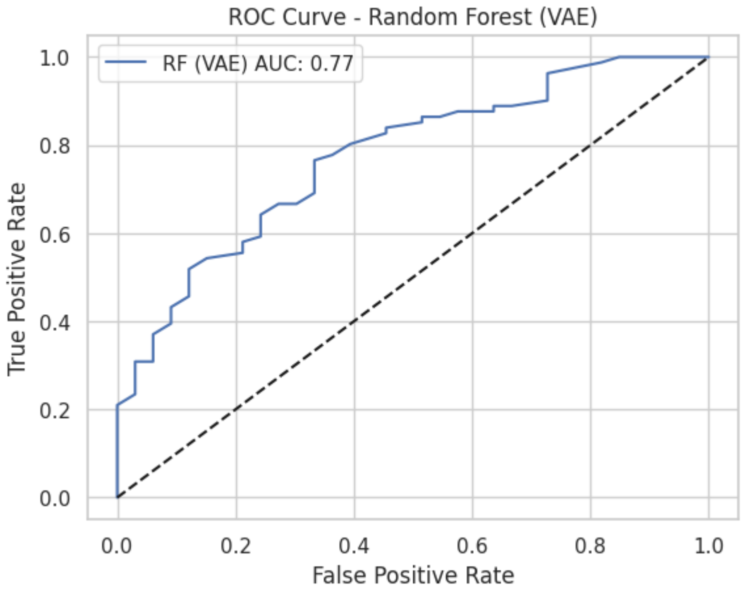


Figure 19. ROC Curves for Random Forest (VAE)

b. ROC Curves for Logistic Regression

Figureure 20 (No Balancing, AUC ≈ 0.82) indicates reasonable performance but with bias. Figureure 21 (SMOTETomek, AUC ≈ 0.83) demonstrates enhanced performance with balanced class detection, while Figureure 22 (VAE, AUC ≈ 0.81) shows a marginal improvement, aligning with its balanced recall of 0.70 for both classes.

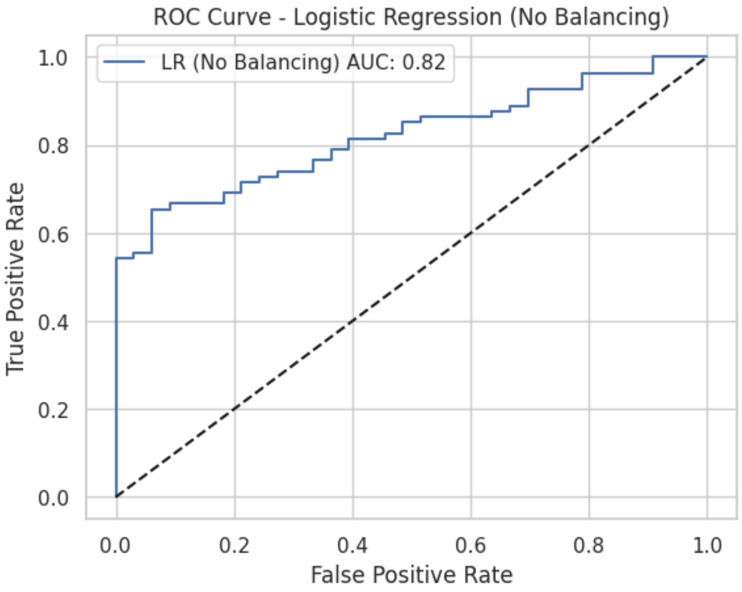


Figure 20. ROC Curves for Logistic Regression (No Balancing)

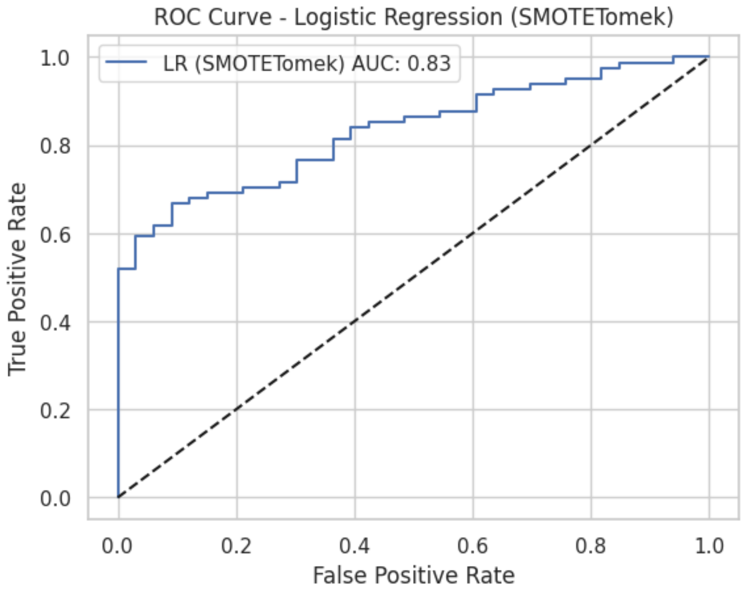


Figure 21. ROC Curves for Logistic Regression (SMOTE)

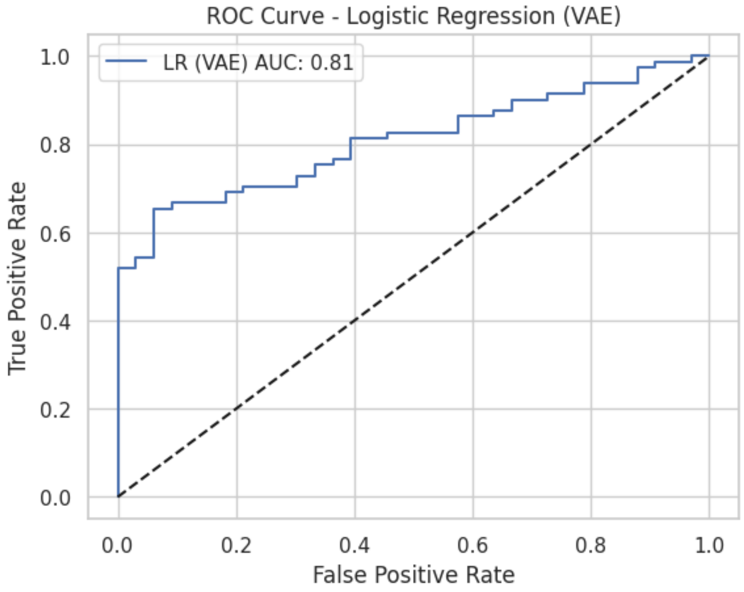


Figure 22. ROC Curves for Logistic Regression (VAE)

c.ROCCurvesforDecisionTree

Figureure 23 (No Balancing, AUC ≈ 0.51) reflects the model's poor performance due to imbalance. Figureure 24 (SMOTETomek, AUC ≈ 0.55) and Figureure 25 (VAE, AUC ≈ 0.54) show slight improvements, but the Decision Tree remains the weakest performer, consistent with its low F1 macro scores.

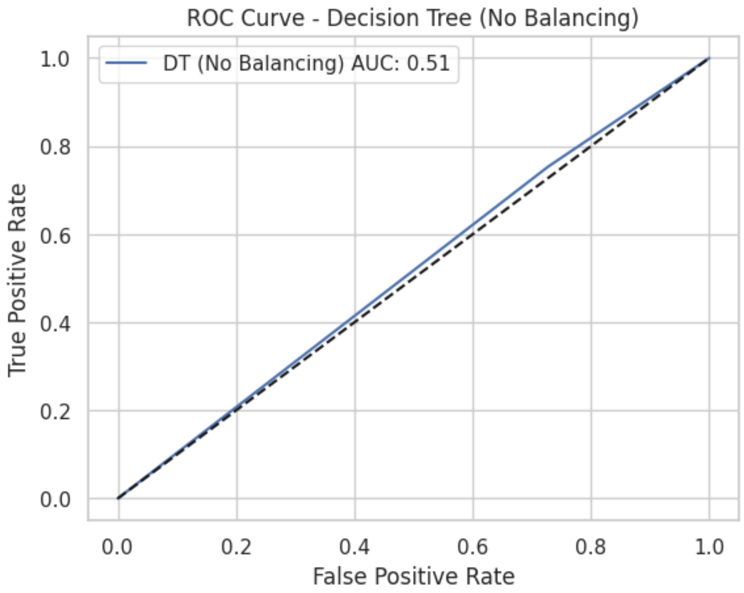


Figure 23. ROC Curves for Decision Tree (No Balancing)

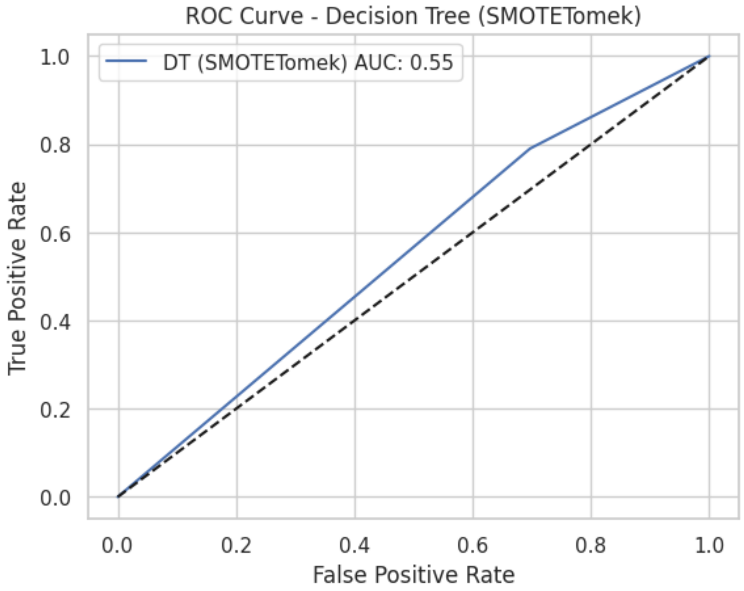


Figure 24. ROC Curves for Decision Tree (SMOTE)

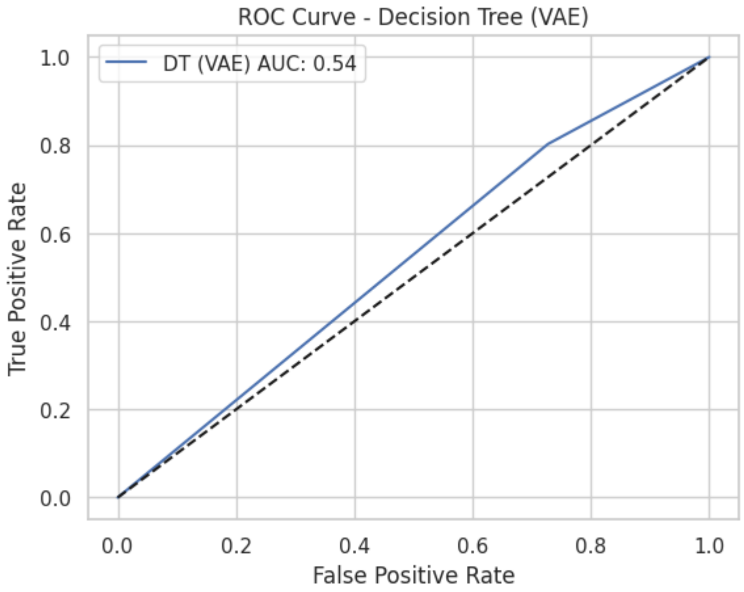


Figure 25. ROC Curves for Decision Tree (VAE)

d. ROC Curves for SVM

Figureure 26 (No Balancing, AUC ≈ 0.70) is heavily skewed toward the majority class (Class 1 recall: 1.00). Figureure 27 (SMOTETomek, AUC ≈ 0.82) shows a significant improvement, aligning with its high F1 macro (0.71). Figureure 28 (VAE, AUC ≈ 0.79) performs better than no balancing but lags behind SMOTETomek.

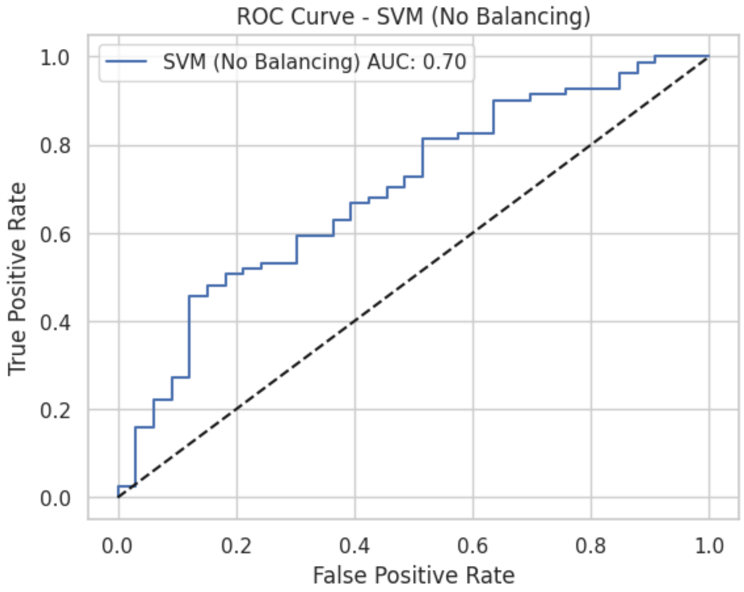


Figure 26. ROC Curves for SVM (No Balancing)

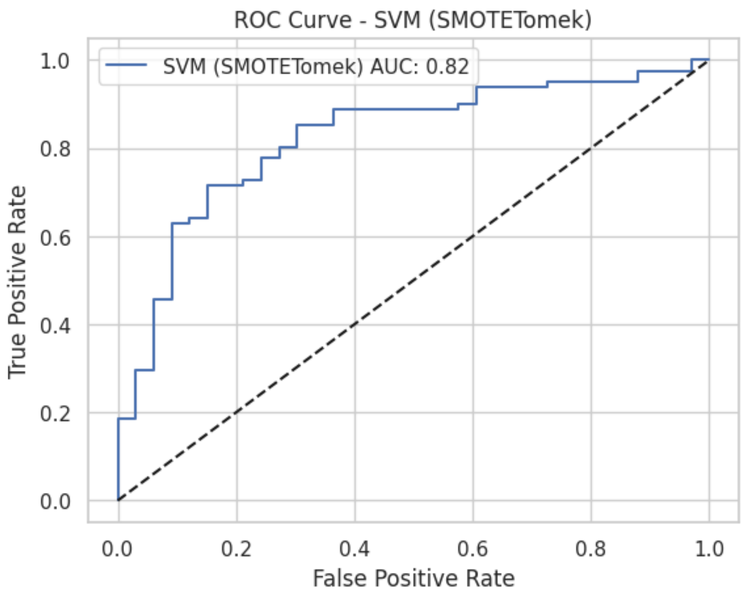


Figure 27. ROC Curves for SVM (SMOTE)

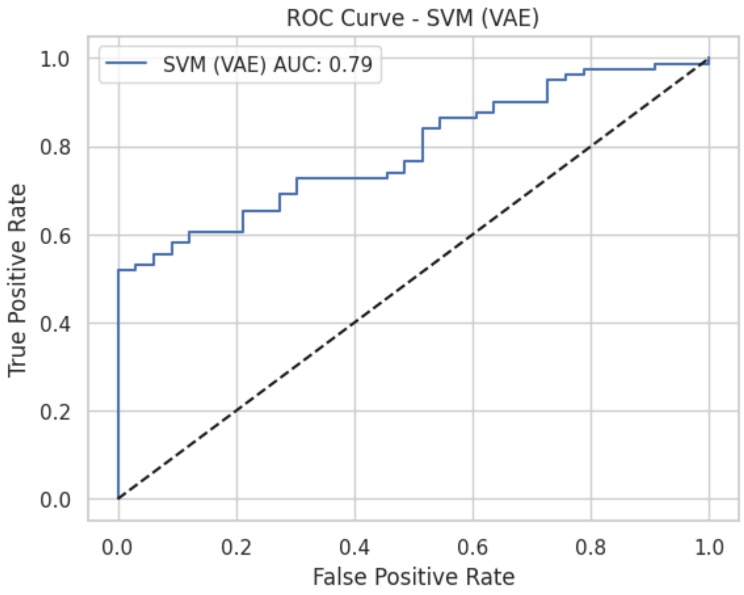


Figure 28. ROC Curves for SVM (VAE)

f. ROC Curves for Neural Network
Figureure 29 (No Balancing, AUC ≈ 0.77) indicates poor minority class detection (Class 0 recall: 0.00). Figureure 30 (SMOTETomek, AUC ≈ 0.82) achieves the highest AUC, reflecting its balanced performance (F1 macro: 0.73). Figureure 31 (VAE, AUC ≈ 0.83

g.) improves over no balancing but is less effective than SMOTETomek, consistent with its lower F1 macro (0.65).

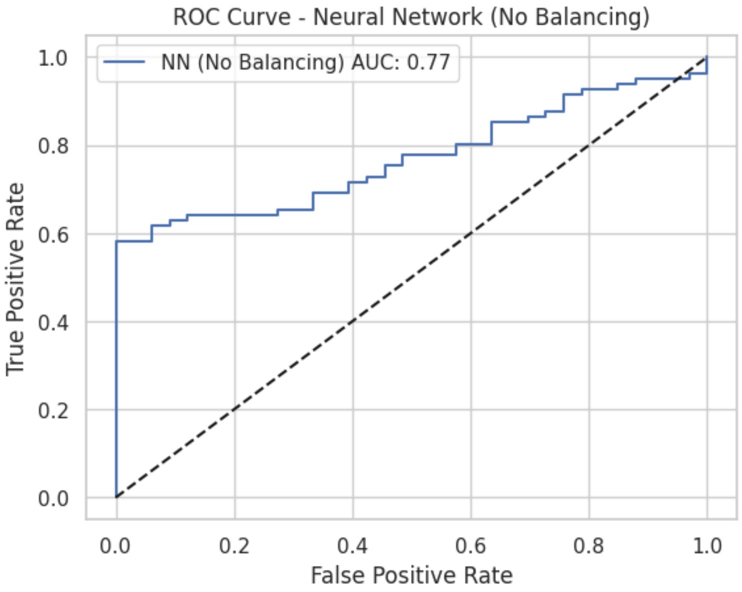


Figure 29. ROC Curves for Neural Network (No Balancing)

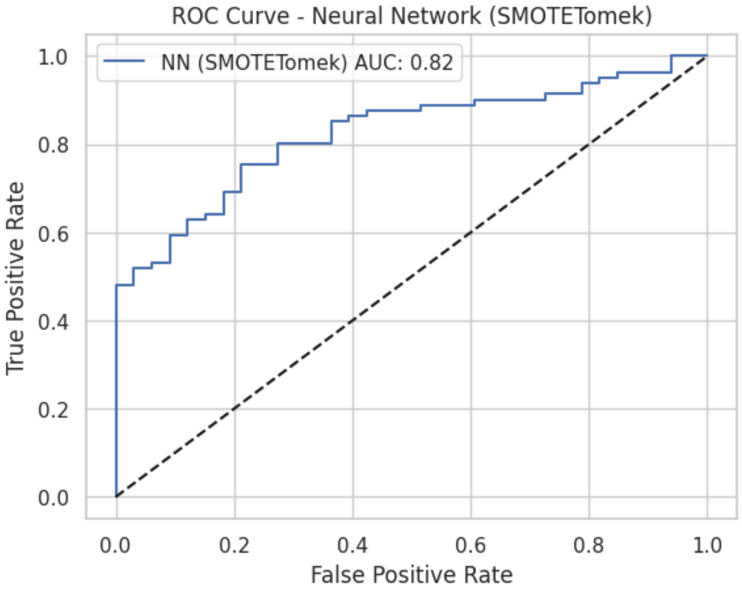


Figure 30. ROC Curves for Neural Network (SMOTE)

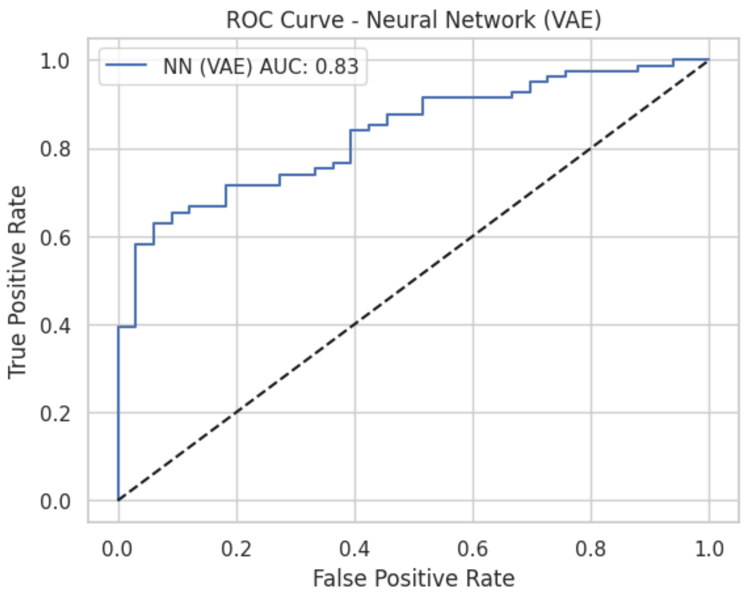


Figure 31. ROC Curves for Neural Network (VAE)

The ROC curves confirm that SMOTETomek consistently enhances model performance across all algorithms, with the Neural Network achieving the highest AUC (0.76), underscoring its superior discriminative ability when class imbalance is addressed.

The evaluation demonstrates that biochemical markers like bilirubin and liver enzymes are the most predictive features for liver disease, as identified through EDA and reinforced by model performance. Preprocessing and feature engineering, including log transformations and outlier capping, enhanced feature utility. SMOTETomek balancing proved most effective, with the Neural Network achieving the best performance (75% accuracy, 0.73 F1 macro), offering balanced classification across both classes. These results highlight the importance of addressing class imbalance and leveraging advanced techniques for accurate liver disease prediction, aligning with prior research [1], [6] [10].

3.5 DISCUSSION

The liver disease prediction project using the Indian Liver Patient Dataset (ILPD) yields several insights into the application of machine learning for medical diagnostics, particularly in addressing class imbalance and identifying predictive biomarkers. The exploratory data analysis (EDA) confirmed that log-transformed biochemical markers, such as log_Direct_Bilirubin (correlation: 0.31) and log_Aspartate_Aminotransferase (0.30), were the strongest predictors of liver disease, aligning with clinical understanding of liver function tests

[1]. However, the limited predictive power of features like Total_Proteins (correlation: -0.04) suggests that protein-related markers may not be as effective for binary classification in this context, potentially due to their overlapping distributions across diseased and non-diseased groups, as observed in swarm plots during EDA. The class imbalance (71.36% diseased, 28.64% non-diseased) significantly impacted model performance when no balancing was applied, with models like SVM and Neural Network achieving near-perfect recall for the majority class (1.00 and 0.99, respectively) but failing entirely on the minority class (0.00 recall for Class 0). This bias underscores the necessity of balancing techniques in medical datasets, where misclassification of the minority class can have severe consequences, such as delayed diagnosis [6].

The implementation of SMOTETomek and Variational Autoencoder (VAE) balancing techniques markedly improved model fairness. SMOTETomek proved superior, enabling the Neural Network to achieve the best performance (75% accuracy, 0.73 F1 macro, AUC 0.76) with balanced recalls (0.79 for Class 0, 0.74 for Class 1), as evidenced by ROC curve analysis. VAE, while effective in generating synthetic minority samples, resulted in a lower F1 macro score (0.65 for Neural Network), possibly due to the challenge of capturing the complex distribution of non-diseased patients with a latent space of 8 units. The choice of a Neural Network with a 64-32-1 architecture, incorporating dropout layers (0.3 and 0.2 rates) and early stopping (patience=30), was effective in preventing overfitting, but its performance on the minority class still lagged slightly behind the majority class, indicating room for further optimization. The gender imbalance in the ILPD (75.64% male) raises concerns about potential biases in model predictions across genders, as the dataset may not fully represent female patients' clinical profiles, a limitation that warrants further investigation in future studies [2]. Overall, the project demonstrates the potential of machine learning to support non-invasive liver disease diagnosis, but it also highlights the challenges of working with imbalanced and demographically skewed medical datasets.

4. Chapter 4 CONCLUSION

This project successfully developed a machine learning framework to predict liver disease using the ILPD, addressing key challenges like class imbalance and feature skewness. EDA identified biochemical markers, particularly log-transformed bilirubin and liver enzymes, as the most predictive features, with correlations up to 0.31, while protein-related features showed limited utility. Preprocessing steps, including imputation of missing values (e.g., Albumin_and_Globulin_Ratio mean: 0.947), log transformations, and outlier capping, prepared the dataset for modeling. SMOTETomek balancing outperformed VAE, enabling the Neural Network to achieve the best results: 75% accuracy, 0.73 F1 macro, and an AUC of 0.76, with balanced recalls (0.79 for Class 0, 0.74 for Class 1). ROC curve analysis further validated SMOTETomek's effectiveness in improving discriminative ability across all models [1], [6]. The project underscores the importance of addressing class imbalance in medical datasets and highlights the potential of machine learning for early liver disease detection, offering a non-invasive tool to support clinical decision-making.

4.1 FUTURE WORK

Future work can enhance the model's applicability and performance in several ways. First, expanding the dataset to include a more diverse population, particularly increasing female representation (currently 24.36% in ILPD), would reduce gender bias and improve generalizability across demographics. Second, incorporating additional features, such as lifestyle factors (e.g., alcohol consumption, smoking history) or genetic markers, could enhance predictive power, as these are known risk factors for liver disease. Third, exploring advanced balancing techniques, such as Generative Adversarial Networks (GANs), may generate more realistic synthetic samples than VAE, potentially improving minority class detection further. Fourth, deploying the model in a real-world clinical setting requires validation with prospective patient data and integration into electronic health record systems, ensuring compliance with regulatory standards like the CDSCO in India. Finally, developing an explainability framework, such as SHAP (SHapley Additive exPlanations), would provide clinicians with insights into feature contributions (e.g., why log_Direct_Bilirubin is highly predictive), fostering trust and adoption in medical practice [10].

4.2 REFLECTION

Reflecting on this project, the process of building a liver disease prediction model revealed both strengths and challenges in applying machine learning to medical data. The use of SMOTETomek to address class imbalance was a significant success, as it enabled balanced performance across classes, a critical requirement in healthcare applications where false negatives can be detrimental. However, the VAE's underperformance (F1 macro: 0.65 vs. 0.73 with SMOTETomek) suggests that its architecture (16-unit hidden layer, 8-unit latent space) may need further tuning to capture the minority class distribution effectively. The project also highlighted the importance of EDA in identifying key features, but the limited predictive power of some features (e.g., Total_Proteins) indicates that feature selection could be refined, perhaps using techniques like recursive feature elimination. Working with the ILPD exposed the challenge of demographic bias, prompting a deeper consideration of fairness in model development. Overall, the project enhanced my understanding of imbalanced datasets and the practical considerations of deploying AI in healthcare, emphasizing the need for collaboration with domain experts to ensure clinical relevance.

REFERENCES

[1] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Informatics in Medicine Unlocked*, vol. 30, p. 101155, 2022., DOI: 10.1016/j.imu.2022.101155,

URL: https://www.sciencedirect.com/science/article/pii/S2352914822002921

[2] K. Gupta, N. Jiwani, N. Afreen, and D. Divyarani, "Liver Disease Prediction using Machine learning Classification Techniques," in *Proc. 2022 IEEE 11th Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, 2022, pp. 221-226., DOI: 10.1109/CSNT54456.2022.9787574,

URL: https://ieeexplore.ieee.org/abstract/document/9787574

[3] C.-C. Wu et al., "Prediction of fatty liver disease using machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 170, pp. 23-29, 2019., DOI: 10.1016/j.cmpb.2018.12.032,

URL: https://www.sciencedirect.com/science/article/abs/pii/S0169260718315724

[4] S. Sontakke, J. Lohokare, and R. Dani, "Diagnosis of liver diseases using machine learning," in *Proc. 2017 Int. Conf. Emerging Trends Innov. ICT (ICEI)*, 2017, pp. 129-133., DOI: 10.1109/ETIICT.2017.7977023,

URL: https://ieeexplore.ieee.org/abstract/document/7977023

[5] T. M. Ghazal et al., "Intelligent Model to Predict Early Liver Disease using Machine Learning Technique," in *Proc. 2022 Int. Conf. Bus. Anal. Technol. Secur. (ICBATS)*, 2022, pp. 1-5., DOI: 10.1109/ICBATS54253.2022.9758929,

URL: https://ieeexplore.ieee.org/abstract/document/9758929

[6] A. Spann et al., "Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review," *Hepatology*, vol. 71, no. 3, pp. 1093-1105, 2020., DOI: 10.1002/hep.31103,

URL: https://aasldpubs.onlinelibrary.wiley.com/doi/pdf/10.1002/hep.31103

[7] V. J. Gogi and V. M. N., "Prognosis of Liver Disease: Using Machine Learning Algorithms," in *Proc. 2018 Int. Conf. Recent Innov. Elect. Electron. Commun. Eng. (ICRIEECE)*, 2018, pp. 1-5., DOI: 10.1109/ICRIEECE44171.2018.9008482,

URL: https://ieeexplore.ieee.org/abstract/document/9008482

[8] K. Drożdż et al., "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," *Cardiovasc. Diabetol.*, vol. 21, no. 1, p. 240, 2022., DOI: 10.1186/s12933-022-01672-9,

URL: https://link.springer.com/article/10.1186/s12933-022-01672-9

[9] A. Sivasangari, B. J. K. Reddy, A. Kiran, and P. Ajitha, "Diagnosis of Liver Disease using Machine Learning Models," in *Proc. 2020 Fourth Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud) (I-SMAC)*, 2020, pp. 574-578., DOI: 10.1109/I-SMAC49090.2020.9243375,

URL: https://ieeexplore.ieee.org/abstract/document/9243375

[10] M. F. Rabbi, S. M. M. Hasan, A. I. Champa, M. A. Zaman, and M. K. Hasan, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," in *Proc. 2020 2nd Int. Conf. Adv. Inf. Commun. Technol. (ICAICT)*, 2020, pp. 111-116., DOI: 10.1109/ICAICT51780.2020.9333528,

URL: https://ieeexplore.ieee.org/abstract/document/9333528

APPENDICES

Appendix A: Project Proposal

Title : Reinforcing Synthetic Data for Liver Cirrhosis Prediction

1. Introduction Liver cirrhosis is a progressive disease requiring early detection to improve survival rates. Traditional machine learning models struggle with class imbalance and limited medical data, reducing predictive accuracy. This research enhances synthetic data using GANs, VAEs, and SMOTE to develop a more reliable predictive model.

2. Objectives

- Develop a liver cirrhosis prediction model using real and synthetic data.
- Enhance model generalization with SMOTE and VAEs for synthetic data generation.
- Use SMOTE to balance class distribution and improve minority class representation.
- Train and compare RF, SVM, and NN models for optimal performance.
- Evaluate Accuracy, Precision, Recall, and AUC-ROC to determine the best model.
- Ensure ethical AI practices, including data privacy compliance and explainable models for clinical transparency.

3. Methodology

- Data preprocessing: deal with variability, normalize characteristics, and handle missing values. Synthetic Data Generation: To balance the dataset, use SMOTE, VAEs, and GANs.
- Model Training: Use both synthetic and real data to train SVM, NN, and RF models.
- Assess performance by measuring F1-score, accuracy, precision, recall, and AUC-ROC.

4. Expected Outcomes

- Increased model accuracy by augmenting it using fake data.
- Enhanced liver cirrhosis early identification for prompt medical treatment.

5. Conclusion

By using synthetic augmentation to solve data restrictions, this research seeks to improve liver cirrhosis prediction. We will produce a balanced dataset that enhances predictive performance by utilizing VAEs, and SMOTE. The results of this study could have a big influence on patient treatment, clinical judgment, and early diagnosis

.Appendix B: Project Management

Trello was used as a project management tool during the research to guarantee effective task management and project tracking. To divide the workflow into discrete stages, such as data preprocessing, model creation, evaluation, and report writing, a Kanban-style board was made. Columns like "To Do," "In Progress," and "Completed" were used to group tasks, and each one was given a checklist and a clear date. This method guaranteed the timely completion of every milestone, enhanced time management, and allowed for systematic progress tracking. Trello streamlined the development process and preserved project structure by acting as a central location to store pertinent code snippets, datasets, and documentation.

Appendix C: Artefact/Dataset

https://github.com/Sunnyanai-alt/LIVER-CIROHSIS-FINAL

Appendix D: Screencast

https://youtu.be/Bc8yTO3ttYI