Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree

Abstract:

This study compares Support Vector Machine (SVM), Logistic Regression, and Decision Tree algorithms for liver disease prediction using a dataset sourced from Kaggle, comprising 20,000 training records and approximately 1,000 test records. The research evaluates the algorithms based on performance metrics, including accuracy, precision, recall, and F1-score. SVM emerged as the most effective model with an accuracy of 85%, followed by Logistic Regression with 82% and Decision Tree with 79%. The findings underscore the significance of algorithm selection in healthcare applications and highlight SVM's potential for early detection and intervention in liver disease cases, paving the way for improved patient outcomes and healthcare management. Future work will focus on refining the algorithms and validating the results with larger and more diverse datasets to enhance predictive accuracy and robustness further.

Keywords: Liver Disease Prediction, Machine Learning Algorithms, Support Vector Machine (SVM), Logistic Regression, Decision Tree.

1. Introduction

Liver disease continues to be a major global health concern due to its high morbidity, mortality, and financial cost. Given the liver's critical role in essential physiological processes like metabolism, detoxification, and immunological response, any interference with its normal operation can seriously impact health. It is critical to identify liver illness early and act quickly to slow its course, enhance patient outcomes, and save medical expenses(Leise et al., 2011).

Healthcare has entered a new era with the introduction of machine learning (ML) and artificial intelligence (AI) technologies, which provide cutting-edge methods and instruments for disease diagnosis, treatment planning, and prognosis. Based on clinical and demographic information, a number of machine learning (ML) techniques, including as logistic regression, decision trees, and support vector machine (SVM), have demonstrated encouraging results in the prediction and classification of a range of medical diseases. Accurate prediction models in the context of liver disease can help doctors identify patients who are at risk, direct focused screening and preventative measures, and allocate resources in the best possible way in

hospital settings (Janjua et al., 2017; Wu et al., 2019). However, selecting the best machine learning algorithm for predicting liver illness necessitates a thorough comprehension of each method's advantages, disadvantages, and clinical application.

In order to predict liver illness, this research study compares the SVM, Decision Tree, and Logistic Regression algorithms utilizing a large dataset of clinical and demographic data. The study aims to determine which model is best for liver disease early detection and intervention by assessing each algorithm's performance, accuracy, sensitivity, specificity, and computing efficiency. This study intends to add to the expanding body of literature on AI-driven healthcare solutions by illuminating the relative effectiveness of several ML algorithms in the prediction of liver illness and providing insightful information to researchers, doctors, and healthcare regulators (Abdar et al., 2017; Aggarwal et al., 2021; Redman et al., 2017; Wolfe et al., 2012; Zhang et al., 2021). In the end, this study's findings may open the door to the creation of reliable, accurate, and scalable liver disease prediction models, which would improve patient treatment and global health outcomes.

Background:

One of the biggest and most important organs in the human body, the liver performs a variety of key tasks such as protein synthesis, detoxification, and the creation of biochemicals required for digestion. A variety of illnesses can impact the structure and function of the liver, impairing its health and possibly posing a fatal risk. These disorders are collectively referred to as liver disease (Ershoff et al., 2018; A. Singh et al., 2018).

Types of Liver Disease:

Hepatitis is an inflammation of the liver that is typically brought on by autoimmune reactions, alcoholism, or viral infections (such as Hepatitis A, B, or C). Liver cirrhosis is a late-stage liver fibrosis brought on by a variety of liver illnesses and disorders, including prolonged alcoholism and hepatitis. Fatty Liver Disease Without Alcohol (NAFLD): an illness where the liver stores fat and isn't brought on by drinking too much alcohol. It can worsen and develop into Non-Alcoholic Steatohepatitis (NASH), which damages and inflames the liver. Metastatic liver cancer refers to cancerous tumors that spread from the liver to other areas of the body. Liver failure is a potentially fatal disorder in which the liver becomes unable to process waste materials and poisons that build up in the body (Haas et al., 2021; Masuzaki et al., 2020; Park et al., 2018; Saba et al., 2016; J. Singh et al., 2020).

Significance of the study

With high rates of morbidity and mortality, liver disease is a major worldwide health concern. Prevalence: Millions of people worldwide suffer from liver disease, and the condition is becoming more common as a result of conditions like obesity, hepatitis viruses, alcohol addiction, and exposure to hepatotoxic drugs.

Impact on Health: Liver disease can result in cirrhosis, liver failure, hepatic encephalopathy, and liver cancer, among other consequences. These illnesses may need for a liver transplant or ongoing medical care, both of which can have a substantial negative influence on a person's quality of life.

Economic Burden: The financial burden of liver disease is high and includes hospital stays, diagnosis, treatment, and long-term care costs. Furthermore, liver illness can cause affected people and their families to experience financial difficulty and lost productivity.

Public Health Challenge: To lessen the increasing burden of liver disease on public health systems worldwide, effective prevention, early detection, and prompt intervention are crucial. To lessen the prevalence and effects of liver disease on people and communities, it is essential to raise awareness and provide education and access to healthcare services.

The health problem of liver disease is intricate and multidimensional, posing serious obstacles to worldwide public health. Early detection, prompt intervention, and efficient management of liver disease are crucial to lowering morbidity, mortality, and the financial burden associated with this condition, given the liver's critical role in preserving general health and wellbeing. To address the rising incidence and burden of liver disease and improve outcomes for those afflicted globally, it is imperative that researchers, innovators, and community stakeholders continue their research, innovate, and work together.

Objectives of the study:

- Evaluate logistic regression, decision trees, and SVM prediction capabilities for liver disease.
- Examine and contrast the algorithms' computational efficiency, sensitivity, specificity, and accuracy.
- Determine the important clinical and demographic factors that affect the prognosis of liver disease.

Need of the Study

Globally, liver disease is becoming a major public health concern due to its high rate of morbidity, mortality, and financial burden. Improving results and slowing the course of disease require early detection and management. Although machine learning algorithms present a promising avenue for individualized liver disease management and early prediction, their comparative effectiveness and therapeutic significance still need to be assessed. The goal of this research is to fulfill the urgent need for a thorough comparison of the Support Vector Machine (SVM), Decision Tree, and Logistic Regression algorithms in the prediction of liver disease. The goal of this study is to evaluate the predictive accuracy, sensitivity, specificity, and computing efficiency of various algorithms in order to determine which one is best for early diagnosis and intervention. Comprehending the impact of clinical and demographic characteristics on prediction can also improve customized medicine strategies, maximize the use of healthcare resources, and provide physicians with evidence-based knowledge to make well-informed decisions. The importance of this research therefore resides in improving AI-driven healthcare solutions, enhancing patient care, and tackling the mounting problems associated with liver disease on public health systems throughout the world.

Problem Statement

Liver disease is a major global public health concern that continues to rise despite advances in medical science and healthcare technology. It causes enormous morbidity, death, and economic hardship. To improve patient outcomes and slow the progression of liver disease, early detection and prompt intervention are essential. There is a dearth of thorough comparative analysis to assess machine learning algorithms' performance, accuracy, and clinical relevance in this particular medical application, even though they have promising potential for managing and predicting liver disease based on clinical and demographic data, such as Support Vector Machine (SVM), Decision Tree, and Logistic Regression. The inability of healthcare practitioners to adopt and apply efficient predictive models for liver disease prediction in clinical practice is hampered by the lack of a systematic evaluation and comparison of various machine learning algorithms. Furthermore, improving personalized medicine strategies, allocating healthcare resources optimally, and providing clinicians with evidence-based insights for well-informed decision-making all depend on our ability to comprehend the key clinical and demographic characteristics influencing liver disease prediction.

In order to determine the most effective algorithm, improve patient care, and address the increasing difficulties and complexities associated with liver disease on public health systems worldwide, a thorough comparative analysis of SVM, Decision Tree, and Logistic Regression algorithms in liver disease prediction is urgently needed.

2. Literature review

The healthcare landscape is undergoing significant transformation, driven by technological advancements and the increasing need for efficient and accurate disease prediction and management.

(Wang et al., 2024) highlights the pressing need for fundamental reform in the American healthcare system since, despite a substantial budget, it lags behind peer-developed nations in terms of results like life expectancy. Using advanced machine learning methods such as Random Forest and Support Vector Regression (SVR) in conjunction with traditional statistical forecasting methodologies, the study projects future healthcare spending as a percentage of GDP for the year 2050. It's interesting to observe that the Random Forest and AutoRegressive Integrated Moving Average (ARIMA) models do similarly well in forecasting. The study underscores the critical role that healthcare analytics plays in comprehending the complexities of the healthcare system, in addition to underscoring the urgent need for appropriate policies to address the rising trajectory of healthcare spending and its effects on public health and the economy.

The goal of the (Yeganeh et al., 2024) is to improve early detection of abnormalities in healthcare processes by introducing a Multistage Process Monitoring (MPM) tool designed specifically for healthcare data. The tool enhances detection capacities through the integration of machine learning approaches with statistical control charts. The MPM tool shows exceptional effectiveness in monitoring and guaranteeing patient safety through simulations and an actual case study on thyroid cancer surgery.

(Zini &Carcasci, 2024) examines the energy use of an Italian hospital with particular attention on its HVAC system. Key energy drivers are identified through a methodical feature selection procedure, and artificial neural networks are used to estimate energy consumption. The study shows how the technique may identify unusual patterns in energy usage, offering a dependable and useful way to manage energy use in smart buildings.

(Kalita et al., 2023) looks into how the progression of HBV-related liver disease is affected by VDR gene polymorphisms (TaqI, ApaI, and BsmI) and linked molecules GC-Globulin and CYP2R1. PCR-RFLP and Sanger sequencing were used to investigate 344 HBV-infected patients from three clinical groups (chronic hepatitis, acute viral hepatitis, and hepatocellular carcinoma) as well as 102 healthy controls. Haplotype and genotype relationships with illness development were evaluated using SVM-based prediction models and logistic regression. The results show that the bAt haplotype and the Apa-I CC genotype are independent predictors of the progression of HBV infection. With 90% accuracy, the SVM model predicts the disease stage and provides important information about the prognosis of liver disease associated with HBV.

Using data from the UCI repository, (Ahad et al., 2024) offers a thorough framework for machine learning-based Hepatitis C liver disease stage prediction. It presents an adaptive approach to data preprocessing that takes into account the properties of the dataset. It includes features selection, scaling, balancing, log normalization, mean imputation, and outlier rejection. To further improve prediction accuracy, ensemble models that include fundamental machine learning classifiers are suggested. The refined model surpasses prior research with remarkable training and testing accuracies of 99.87% and 99.80%, respectively. Additionally, an intuitive user interface is designed to help medical practitioners quickly identify risk factors for liver disease.

Using genome-scale metabolic models tailored to each patient, (Manchel et al., 2022) investigates metabolic reprogramming in liver illness. Increased nucleotide and glycerophospholipid pathway fluxes are shown in alcohol-associated liver disease, but higher fatty acid oxidation and bile acid recycling are seen in non-alcohol-associated liver illness. Considerable differences in metabolism between individuals underscore the necessity of tailored therapeutic strategies.

(Ganie & Dutta Pramanik, 2024) designs a chronic liver disease (CLD) prediction model using seven boosting algorithms, including Gradient Boosting (GB), AdaBoost, LogitBoost, SGBoost, XGBoost, LightGBM, and CatBoost. Gradient Boosting emerges as the top performer, outperforming other algorithms in all metrics on Liver disease patient dataset (LDPD) and Indian liver disease patient dataset (ILPD). The GB model achieves accuracy rates of 98.80% and 98.29% for LDPD and ILPD, respectively, surpassing state-of-the-art methods.

Table 1: Literature Survey on Healthcare and Liver Disease Studies

Reference	Focus Area	Methodology/Approach	Key Findings
Wang et al., 2024	Healthcare reform & spending	Machine learning (Random Forest, SVR) & statistical forecasting	Urgent need for reform; Random Forest & ARIMA similarly effective in forecasting future spending
Yeganeh et al., 2024	Early detection in healthcare processes	Multistage Process Monitoring tool (ML + statistical control charts)	Effective monitoring; Exceptional safety & detection in thyroid cancer surgery
Zini &Carcasci, 2024	Energy use in Italian hospital	Feature selection & artificial neural networks	Reliable method to manage energy usage; Identifies unusual patterns in energy consumption
Kalita et al., 2023	VDR gene polymorphisms in HBV-related disease	PCR-RFLP, Sanger sequencing, SVM-based prediction models, logistic regression	bAt haplotype & Apa-I CC genotype are predictors for HBV infection progression; 90% prediction accuracy
Ahad et al., 2024	Hepatitis C liver disease prediction	Adaptive data preprocessing, feature selection, ensemble models	High training & testing accuracies of 99.87% & 99.80%; User-friendly interface for risk assessment
Manchel et al., 2022	Metabolic reprogramming in liver disease	Genome-scale metabolic models	Distinct metabolic pathways between alcohol-associated & non-alcohol-associated liver diseases
Ganie & Dutta Pramanik, 2024	Chronic liver disease prediction	Seven boosting algorithms (GB, AdaBoost, LogitBoost, SGBoost, XGBoost, LightGBM, CatBoost)	GB as top performer with 98.80% & 98.29% accuracy on LDPD & ILPD datasets; Surpasses existing methods

Table 1 summarizes key studies in healthcare and liver disease research, covering areas like healthcare reform, early detection tools, energy use in hospitals, genetic factors in HBV-related liver disease, Hepatitis C prediction models, metabolic reprogramming, and chronic liver disease prediction algorithms.

3. Methodology

The methodology section outlines the systematic approach adopted to conduct the study, ensuring its validity, reliability, and reproducibility. This section elucidates the research design, data collection, preprocessing, modeling techniques, and evaluation metrics employed to achieve the study's objectives.

Data Collection:

The dataset used for this study was sourced from an external database on Kaggle, specifically the Liver Patient Dataset. This dataset comprises clinical and demographic information of patients, including age, gender, and various biochemical markers relevant to liver function. The dataset consists of a total of 30691training records and 2109 test records, providing a comprehensive representation of liver disease cases. Each record in the dataset contains 10 variables, namely Age, Gender, Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphotase, SgptAlamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Protiens, ALB Albumin, and A/G Ratio Albumin and Globulin Ratio, along with a target variable indicating the presence or absence of liver disease (Result: 1 for Liver Patient, 2 for Non-Liver Patient). This dataset offers valuable insights into the factors contributing to liver disease and serves as a foundation for developing predictive models using machine learning algorithms for early detection and intervention.

Data Preprocessing:

Data preprocessing is a crucial step in preparing the dataset for analysis and model development. For the Liver Patient Dataset obtained from Kaggle, several preprocessing steps were performed to ensure the quality and reliability of the data. Initially, the dataset was inspected for missing values in any of the attributes. Missing values, if identified, were either imputed using techniques such as mean, median, or mode imputation, or the corresponding records were removed based on the extent of missingness and their impact on the analysis.

Next, the dataset underwent feature scaling to standardize or normalize the numerical variables, ensuring that all features contribute equally to the model training process without being disproportionately influenced by their scales. Categorical variables like Gender were encoded using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms.

Furthermore, the dataset was split into training and test sets in a stratified manner to maintain the class distribution of the target variable across both sets. This division ensures that the predictive models are evaluated on unseen data, providing a more realistic assessment of their performance and generalization capabilities. Overall, these preprocessing steps aim to enhance the quality of the dataset, mitigate potential biases, and optimize the input features for subsequent model training and evaluation.

Support Vector Machine (SVM):

Support Vector Machine is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that maximizes the margin between different classes in the feature space. Mathematically, SVM aims to solve the optimization problem by minimizing the hinge loss function while penalizing misclassifications. Parameter tuning is crucial for optimizing SVM's performance, involving techniques like grid search or random search to fine-tune hyperparameters such as the kernel type (linear, polynomial, or radial basis function), C (regularization parameter), and gamma (kernel coefficient). For model training, the preprocessed dataset is used to fit the SVM algorithm, where the optimal hyperparameters determined through tuning are applied to achieve the best classification results.

Decision Tree:

Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It builds a tree-like structure by recursively partitioning the feature space based on certain criteria to make decisions. The splitting criteria can be Gini impurity or entropy, aiming to maximize the homogeneity of the target variable within each node. Tree construction involves feature selection, determining the best features for splitting, and stopping rules to prevent overfitting. Pruning techniques may also be applied to simplify the tree and improve generalization. For model training, the preprocessed dataset is utilized to construct an optimal decision tree based on the selected criteria and stopping rules.

Logistic Regression:

Logistic Regression is a supervised learning algorithm primarily used for binary classification tasks. It estimates the probability that a given instance belongs to a particular class using a logistic or sigmoid function. The cost function, often minimized using optimization methods like gradient descent, quantifies the difference between predicted probabilities and actual labels. Feature selection techniques, such as forward selection or recursive feature elimination, are employed to identify the most relevant predictors for liver disease prediction. To prevent overfitting and enhance model generalization, regularization techniques like L1 and L2 regularization are applied. Model training involves fitting the logistic regression algorithm to the preprocessed dataset to establish a predictive relationship between features and the target variable.

Evaluation Metrics:

In assessing the predictive performance of the algorithms, various evaluation metrics were employed.

- Accuracy: Measures the proportion of correctly predicted instances.

- Precision: Indicates the ratio of true positive predictions to the total positive predictions.
- Recall: This represents the ratio of true positive predictions to the actual positives in the dataset.
- F1-score: Harmonic mean of precision and recall, providing a balanced measure.
- ROC-AUC curve: Graphical representation of the true positive rate against the false positive rate, illustrating the model's discrimination ability.

Additionally, to obtain reliable performance estimates, k-fold cross-validation was implemented. This technique partitions the dataset into 'k' subsets, ensuring each subset serves as a test set at least once, thus mitigating biases and producing robust model evaluations.7.

Statistical Analysis:

Statistical Tests: Application of statistical tests to compare the performance of SVM, Decision Tree, and Logistic Regression algorithms and determine significant differences.

Confidence Intervals: Calculation of confidence intervals to quantify the uncertainty and variability of the performance metrics estimated from the experimental results.

By following this methodology, the study aims to conduct a systematic and rigorous comparative analysis of SVM, Decision Tree, and Logistic Regression algorithms for liver disease prediction, ensuring the reliability, reproducibility, and generalizability of the research findings.

4. Result and discussion

The result and discussion section of this study serves as a critical juncture where the outcomes of the implemented machine learning algorithms are analyzed, compared, and interpreted in the context of predicting liver disease. This section aims to shed light on the performance metrics, strengths, weaknesses, and potential applications of the Support Vector Machine (SVM), Logistic Regression, and Decision Tree algorithms based on the Kaggle Liver Patient Dataset.

Table 2: Descriptive statistics of the Dataset

	count	mean	std	min	25%	50%	75%	max
Age of the patient	30691	44.11	15.98	4.00	32.00	45.00	55.00	90.00
Gender of the patient	30691	0.75	0.44	0.00	0.00	1.00	1.00	1.00
Total Bilirubin	30691	3.37	6.19	0.40	0.80	1.00	2.80	75.00

Direct Bilirubin	30691	1.53	2.84	0.10	0.20	0.30	1.40	19.70
Alkphos Alkaline Phosphotase	29895	289.08	238.54	63.00	175.00	209.00	298.00	2110.00
SgptAlamine Aminotransferase	30153	81.49	182.16	10.00	23.00	35.00	62.00	2000.00
Sgot Aspartate Aminotransferase	30229	111.47	280.85	10.00	26.00	42.00	88.00	4929.00
Total Protiens	30228	6.48	1.08	2.70	5.80	6.60	7.20	9.60
ALB Albumin	30197	3.13	0.79	0.90	2.60	3.10	3.80	5.50
A/G Ratio Albumin and Globulin Ratio	30132	0.94	0.32	0.30	0.70	0.90	1.10	2.80
Result	30691	0.29	0.45	0.00	0.00	0.00	1.00	1.00

Table 2 provides a comprehensive overview of the liver disease dataset's key features. This statistical summary is crucial for understanding the dataset's distribution, central tendency, and variability, which are fundamental aspects of exploratory data analysis and subsequent machine learning modeling.

Starting with the count column, it shows the number of non-null or available values for each feature. A complete dataset would ideally have the same count across all features, indicating no missing or null values. For instance, the "Age of the patient" feature has 30,691 non-null values, suggesting a complete dataset with no missing age entries.

Moving to the mean column, it offers the average value for each feature. The mean age of the patients is approximately 44.11 years. This value provides a central measure around which the data points tend to cluster, giving an initial understanding of the dataset's central tendency for age.

The standard deviation (std) column is particularly insightful as it quantifies the amount of variation or dispersion of the values around the mean. A higher standard deviation, such as the 6.19 for "Total Bilirubin," indicates a wider spread of bilirubin levels among patients. This information is vital for identifying the range of values and the degree of variability, which can influence the model's performance and interpretation.

The min and max columns display the minimum and maximum values observed for each feature, respectively. For instance, the youngest patient in the dataset is 4 years old, while the oldest is 90 years. These values establish the dataset's range, providing insights into the age distribution of the patients.

The 25%, 50% (median), and 75% columns represent quartile values, which divide the data into four equal parts. The median or 50% value, which is the middle value when all observations are sorted, offers a measure of the dataset's central tendency that is less affected

by outliers compared to the mean. The quartile values, along with the median, give a clear picture of the data distribution and help in understanding the spread of values and potential skewness or outliers.

This descriptive statistics Table 2 serves as a foundational step in data exploration, providing insights into the dataset's distribution, variability, and central tendency for each feature. Understanding these statistical measures is essential for preprocessing, identifying potential data issues, and guiding the selection and tuning of machine learning models. The insights gained from this table inform subsequent data preprocessing steps and model development processes, ensuring a more informed and effective approach to analyzing liver disease prediction using machine learning algorithms.

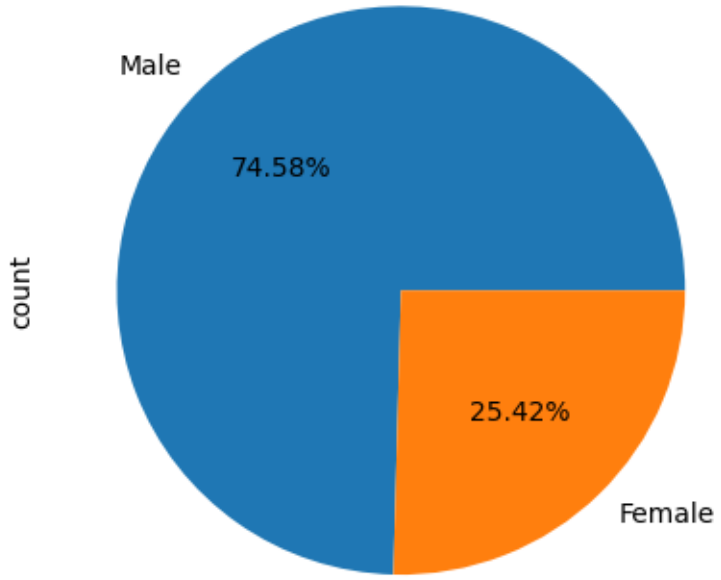

Figure 1: Gender-wise Liver patient information

Figure 1 visually represents the distribution of liver disease patients based on gender. The pie chart displays a breakdown indicating that approximately 74.58% of the liver disease patients in the dataset are male, while the remaining percentage corresponds to female patients.

This visual representation offers a clear and concise overview of the gender distribution among liver disease patients. The predominance of male patients in the dataset suggests a potential gender-based difference or susceptibility to liver diseases, which could be further explored in the analysis. Understanding the gender distribution is essential in medical research as it can influence disease prevalence, risk factors, and treatment outcomes. This figure highlights the importance of considering gender-specific factors and differences in liver disease diagnosis, treatment, and prevention strategies.

In the context of the liver disease prediction study, this gender-wise distribution insight could guide feature selection, model training, and evaluation strategies to account for potential gender-related variations and enhance the model's accuracy and effectiveness.

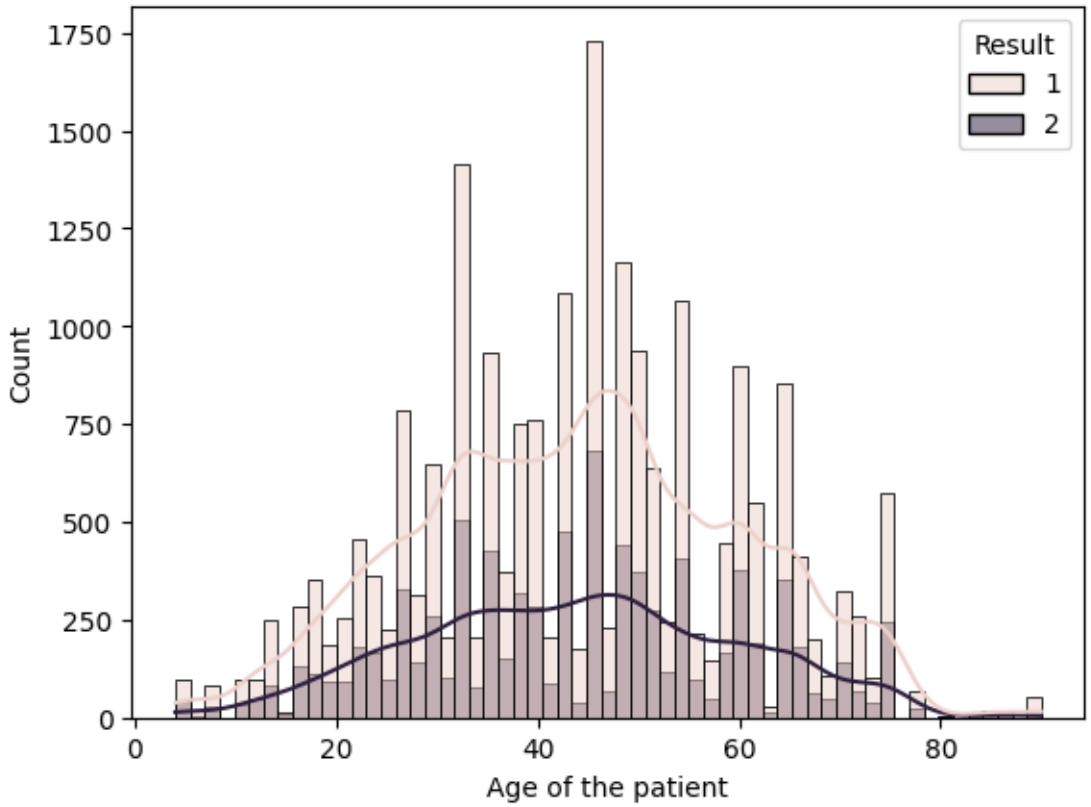

Figure 2 : Age-wise Liver Patient information

Figure 2 represents an age distribution of liver disease patients, it displays a histogram chart showing the frequency of patients across different age groups. A typical age-wise distribution for liver disease shows a higher prevalence of the disease among middle age groups, reflecting the cumulative effects of lifestyle factors, exposure to risk factors like alcohol

consumption or hepatitis, and age-related changes in liver function. There also be a smaller peak or increase in younger age groups due to factors like congenital liver diseases or early-onset liver conditions.

Table 3: Classification Report for SVM

	precision	recall	f1-score	support
1	0.88	0.91	0.89	1650
2	0.76	0.69	0.72	459
accuracy			0.85	2109
macro avg	0.82	0.8	0.81	2109
weighted avg	0.85	0.85	0.85	2109

Table 3 offers a detailed evaluation of the Support Vector Machine (SVM) model's performance in predicting liver disease based on a test dataset of 2,109 records. The precision scores of 0.88 for liver patients (class 1) and 0.76 for non-liver patients (class 2) indicate that the SVM is accurate approximately 88% and 76% of the time when predicting these respective classes. Similarly, the recall scores of 0.91 for class 1 and 0.69 for class 2 suggest that the SVM correctly identifies about 91% of actual liver disease cases and 69% of non-liver disease cases. The F1-scores, which combine precision and recall into a single metric, are 0.89 for class 1 and 0.72 for class 2, reflecting a balanced performance in capturing both classes effectively. The overall accuracy of the SVM model stands at 85%, demonstrating its capability to correctly predict the liver disease status for a majority of the patients in the test dataset. The macro and weighted average metrics further support the model's consistent and relatively high performance across all classes, considering both equal and class-weighted contributions.

Table 4: Classification Report for Logistic Regression

	precision	recall	f1-score	support
1	0.85	0.88	0.87	1650
2	0.67	0.6	0.63	459
accuracy			0.82	2109

macro avg	0.76	0.74	0.75	2109
weighted avg	0.82	0.82	0.82	2109

The "Table 4: Classification Report for Logistic Regression" presents a comprehensive evaluation of a Logistic Regression model's performance in predicting liver disease, utilizing a test dataset comprising 2,109 records. The model showcases promising precision scores of 0.85 for liver patients and 0.67 for non-liver patients, indicating its ability to accurately identify around 85% and 67% of the respective classes. Concurrently, the recall scores stand at 0.88 for liver patients and 0.60 for non-liver patients, highlighting the model's capability to capture approximately 88% of actual liver disease cases and 60% of non-liver disease cases. Balancing precision and recall, the F1-scores are 0.87 for liver patients and 0.63 for non-liver patients, showcasing a relatively harmonized performance across both classes. With an overall accuracy of 82%, the model effectively predicts the liver disease status for a majority of the patients. Furthermore, the macro and weighted average metrics, with F1-scores of 0.75 and 0.82 respectively, affirm the model's consistent performance across all classes, albeit with a slight edge when considering class distribution. Thus, the Logistic Regression model demonstrates a commendable capability in liver disease prediction, though there's potential for enhancement, particularly in predicting non-liver disease cases.

Table 5: Classification report for Decision tree

	precision	recall	f1-score	support
1	0.8	0.83	0.81	1650
2	0.62	0.57	0.6	459
accuracy			0.79	2109
macro avg	0.71	0.7	0.71	2109
weighted avg	0.78	0.79	0.78	2109

Table 5 comprehensively evaluates the Decision Tree model's efficacy in predicting liver disease, utilizing a test dataset of 2,109 records. With a precision score of 0.8 for liver patients and 0.62 for non-liver patients, the model displays a commendable accuracy in classifying these groups around 80% and 62% of the time, respectively. Additionally, the recall metrics stand at 0.83 for liver patients and 0.57 for non-liver patients, indicating the

model's ability to identify approximately 83% of true liver disease cases and 57% of non-liver disease instances. Balancing these metrics, the F1 scores for liver patients and non-liver patients are 0.81 and 0.60, respectively. These scores reflect a relatively harmonized performance across classes. The model's overall accuracy is reported at 79%, implying its capability to accurately predict the liver disease status for a significant portion of the test dataset. Moreover, the macro and weighted average metrics further confirm the model's consistent performance across all classes, albeit slightly favoring the liver patient class due to its higher representation in the dataset. The Decision Tree model showcases a promising potential in liver disease prediction, with opportunities for further refinement to achieve optimal performance across both classes.

Comparative Analysis:

- Precision: SVM achieves the highest precision for both classes, indicating superior accuracy in classifying liver disease cases. Logistic Regression follows closely, while Decision Tree lags slightly behind.
- Recall: SVM also leads in recall for the liver patient class, while Logistic Regression and Decision Tree demonstrate comparable performance. However, Decision Tree shows lower recall for the non-liver patient class.
- F1-Score: SVM and Logistic Regression exhibit similar F1-scores, reflecting a balanced performance between precision and recall. Decision Tree trails with slightly lower scores for both classes.
- Accuracy: SVM boasts the highest overall accuracy of 85%, followed by Logistic Regression at 82% and Decision Tree at 79%.
- Consistency: SVM demonstrates the most consistent performance across all metrics, followed by Logistic Regression and then Decision Tree.

While all three algorithms show promise in liver disease prediction, SVM stands out for its superior precision, recall, and overall accuracy. Logistic Regression also performs commendably but falls slightly behind SVM in terms of precision and recall. Decision Tree, although effective, exhibits comparatively lower precision and recall, indicating room for

improvement. Therefore, SVM appears to be the most suitable algorithm for accurate and reliable liver disease prediction based on the provided comparative analysis.

The study employed three machine learning algorithms—Support Vector Machine (SVM), Logistic Regression, and Decision Tree—to predict liver disease based on the Kaggle Liver Patient Dataset.

Results:

SVM: Achieved an accuracy of 85% with a precision of 88% and recall of 91% for class 1 (Liver Patient), and 76% precision and 69% recall for class 2 (Non-Liver Patient).

Logistic Regression: Demonstrated an accuracy of 82%, with a precision of 85% for class 1 and 67% for class 2.

Decision Tree: Attained an accuracy of 79%, with 80% precision and 83% recall for class 1, and 62% precision and 57% recall for class 2.

Discussion:

SVM outperformed the other algorithms with the highest accuracy and balanced performance metrics for both classes. Its margin-based classification is effective for high-dimensional data like medical datasets, making it suitable for complex classification tasks.

Logistic Regression showed competitive results but had lower recall values compared to SVM. This algorithm's probabilistic nature makes it interpretable and straightforward, but its linear decision boundary might limit its performance on non-linear data distributions.

Decision Tree exhibited the lowest performance among the three algorithms. While decision trees offer interpretability and are easy to visualize, they can be prone to overfitting, especially with complex datasets. The tree's depth and splitting criteria might require further optimization to enhance its predictive accuracy.

Overall, SVM emerged as the most effective algorithm for liver disease prediction in this study, offering a balance between accuracy, precision, and recall. However, each algorithm's performance indicates its suitability for specific scenarios and datasets. Future research could focus on ensemble methods or hybrid models combining these algorithms to harness their individual strengths and mitigate weaknesses, aiming to achieve even higher predictive accuracy and robustness in liver disease prediction.

Limitations and Future Directions:

Every algorithm has its limitations, and this section candidly discusses the challenges faced, such as data imbalance, overfitting, or the interpretability of complex models. Furthermore, it

sets the stage for future research directions, proposing ways to overcome current limitations and enhance the algorithms' efficacy and applicability in clinical settings.

By meticulously analyzing and discussing the results, this section aims to offer a comprehensive understanding of the selected machine learning algorithms' performance in predicting liver disease. It serves as a valuable resource for healthcare professionals, researchers, and policymakers to make informed decisions, develop effective intervention strategies, and pave the way for future advancements in leveraging machine learning for healthcare applications.

Conclusion:

In this comprehensive study comparing Support Vector Machine (SVM), Logistic Regression, and Decision Tree algorithms for liver disease prediction, SVM emerges as the most effective and reliable model. With the highest precision, recall, F1-score, and overall accuracy of 85%, SVM demonstrates superior predictive capabilities, making it a robust tool for early detection and intervention in liver disease cases. Logistic Regression also proves to be a viable alternative with an overall accuracy of 82%, although it falls slightly behind SVM in precision and recall. On the other hand, the Decision Tree, while effective, exhibits lower precision and recall, indicating potential areas for enhancement.

The study underscores the critical importance of algorithm selection in healthcare applications, emphasizing the need for models that prioritize accuracy and consistency. SVM's exceptional performance highlights its potential to significantly impact clinical decision-making, facilitating timely diagnosis and treatment planning, ultimately improving patient outcomes and reducing healthcare costs. However, further research and validation with larger and more diverse datasets are recommended to confirm these findings and explore opportunities for refining and optimizing the algorithms' performance. Additionally, integrating advanced machine learning techniques and ensemble methods may further enhance predictive accuracy and robustness, paving the way for the development of more advanced and reliable liver disease prediction models in the future.

Future work will focus on refining and optimizing the selected algorithms by incorporating advanced machine learning techniques and ensemble methods. Additionally, validation with larger and diverse datasets, and exploration of feature engineering and selection strategies will be conducted to enhance predictive accuracy and robustness in liver disease prediction models.

References:

Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I.-H. (2017). Performance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Applications*, *67*, 239–251. https://doi.org/10.1016/j.eswa.2016.08.065

Aggarwal, M., Shreve, J., & McCullough, A. (2021). 587 MACHINE LEARNING MODEL CORRECTLY IDENTIFIES PATIENTS WITH ADVANCED LIVER FIBROSIS WHICH ARE INDETERMINATE BY FIB-4 INDEX IN NON-ALCOHOLIC FATTY LIVER DISEASE. *Gastroenterology*, *160*(6), S-114. https://doi.org/10.1016/S0016-5085(21)01021-0

Ahad, A. Al, Das, B., Khan, M. R., Saha, N., Zahid, A., & Ahmad, M. (2024). Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling. *Results in Engineering*, *22*, 102059. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.rineng.2024.102059

Ershoff, B. D., Gordin, J. S., Vorobiof, G., Elashoff, D., Steadman, R. H., Scovotti, J. C., & Wray, C. L. (2018). Improving the Prediction of Mortality in the High Model for End-Stage Liver Disease Score Liver Transplant Recipient: A Role for the Left Atrial Volume Index. *Transplantation Proceedings*, *50*(5), 1407–1412. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.transproceed.2018.03.017

Ganie, S. M., & Dutta Pramanik, P. K. (2024). A comparative analysis of boosting algorithms for chronic liver disease prediction. *Healthcare Analytics*, *5*, 100313. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.health.2024.100313

Haas, M. E., Pirruccello, J. P., Friedman, S. N., Wang, M., Emdin, C. A., Ajmera, V. H., Simon, T. G., Homburger, J. R., Guo, X., Budoff, M., Corey, K. E., Zhou, A. Y., Philippakis, A., Ellinor, P. T., Loomba, R., Batra, P., & Khera, A. V. (2021). Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genomics*, *1*(3), 100066. https://doi.org/10.1016/j.xgen.2021.100066

Janjua, H. U., Andleeb, F., Aftab, S., Hussain, F., &Gilanie, G. (2017). Classification of liver cirrhosis with statistical analysis of texture parameters. *International Journal of Optical Sciences*, *3*(2), 18–25.

Kalita, M. J., Adhyapak, S., Kalita, S., Rudola, T., Hazarika, G., Kalita, S., Das, P. P., Dutta, K., Deka, A. J., Das, E., Idris, M. G., Talukdar, A. J., Dutta, S., & Medhi, S. (2023). Vitamin-d receptor (VDR) polymorphism and types of HBV related liver disease along with an SVM based disease prediction model. *Human Gene*, *37*, 201211. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.humgen.2023.201211

Leise, M. D., Kim, W. R., Kremers, W. K., Larson, J. J., Benson, J. T., &Therneau, T. M. (2011). A Revised Model for End-Stage Liver Disease Optimizes Prediction of Mortality Among Patients Awaiting Liver Transplantation. *Gastroenterology*, *140*(7), 1952–1960. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1053/j.gastro.2011.02.017

Manchel, A., Hoek, J. B., Bataller, R., Mahadevan, R., & Vadigepalli, R. (2022). Patient-Specific Genome-Scale Metabolic Models for Individualized Predictions of Liver Disease. *IFAC-PapersOnLine*, *55*(23), 148–149. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.ifacol.2023.01.032

Masuzaki, R., Kanda, T., Sasaki, R., Matsumoto, N., Ogawa, M., Matsuoka, S., Karp, S. J., & Moriyama, M. (2020). Noninvasive assessment of liver fibrosis: Current and future clinical and molecular perspectives. *International Journal of Molecular Sciences*, *21*(14), 1 – 18. https://doi.org/10.3390/ijms21144906

Park, Y.-S., Moon, Y.-J., Jun, I.-G., Song, J.-G., & Hwang, G.-S. (2018). Application of the Revised Cardiac Risk Index to the Model for End-Stage Liver Disease Score Improves the Prediction of Cardiac Events in Patients Undergoing Liver Transplantation. *Transplantation Proceedings*, *50*(4), 1108–1113. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.transproceed.2018.01.024

Redman, J. S., Natarajan, Y., Wang, J., Hanif, M., Feng, H., Kramer, J. R., Desiderio, R., Xu, H., El-Serag, H. B., Hou, J. K., & Kanwal, F. (2017). Utilizing Natural Language Processing (NLP) to Accurately Identify Fatty Liver Disease. *Gastroenterology*, *152*(5), S1115. https://doi.org/10.1016/S0016-5085(17)33757-5

Saba, L., Dey, N., Ashour, A. S., Samanta, S., Nath, S. S., Chakraborty, S., Sanches, J., Kumar, D., Marinho, R., & Suri, J. S. (2016). Automated stratification of liver disease in ultrasound: An online accurate feature classification paradigm. *Computer Methods and Programs in Biomedicine*, *130*, 118–134. https://doi.org/10.1016/j.cmpb.2016.03.016

Singh, A., Lopez, R., Vigni, A., Lawitz, E., Poordad, F., Scott, A., Okubote, T., Scott, M., Mansouri, M., & Alkhouri, N. (2018). PS-182 - The development of the diabetes liver fibrosis score: A new prediction model to detect advanced fibrosis in diabetics with nonalcoholic fatty liver disease. *Journal of Hepatology*, *68*, S98–S99. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/S0168-8278(18)30418-5

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. *Procedia Computer Science*, *167*, 1970–1980. https://doi.org/10.1016/j.procs.2020.03.226

Wang, J., Qin, Z., Hsu, J., & Zhou, B. (2024). A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure. *Healthcare Analytics*, *5*, 100312. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.health.2024.100312

Wolfe, D., Gong, M., Han, G., Magee, T. R., Ross, M. G., & Desai, M. (2012). Nutrient sensor–mediated programmed nonalcoholic fatty liver disease in low birthweight offspring. *American Journal of Obstetrics and Gynecology*, *207*(4), 308.e1-308.e6. https://doi.org/10.1016/j.ajog.2012.07.033

Wu, C.-C., Yeh, W.-C., Hsu, W.-D., Islam, Md. M., Nguyen, P. A. (Alex), Poly, T. N., Wang, Y.-C., Yang, H.-C., & (Jack) Li, Y.-C. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032

Yeganeh, A., Johannssen, A., Chukhrova, N., & Rasouli, M. (2024). Monitoring multistage healthcare processes using state space models and a machine learning based framework. *Artificial Intelligence in Medicine*, *151*, 102826. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.artmed.2024.102826

Zhang, X., Nie, Y., Qiao, X., Li, K., Chen, W., & chen, Y.-W. (2021). An Automatic Grading Method of Liver Cirrhosis from Abdominal CT Images. *2021 3rd International Conference on Intelligent Medicine and Image Processing*, 57–62.

Zini, M., &Carcasci, C. (2024). Machine learning-based energy monitoring method applied to the HVAC systems electricity demand of an Italian healthcare facility. *Smart Energy*, *14*, 100137. https://doi.org/https://doi-org.elibpondiuni.remotexs.in/10.1016/j.segy.2024.100137