Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Machine Learning-Based Predictive Modeling for Early Detection of Liver Cirrhosis

Dr. Jadhav Nitin B[1], Dr. Archana S. Banait[2], Dr. Desai Jabbar V.[3], Dr. Ranjit M. Gawande[4], Dr. Satish V. Kakade[5], Sonal Dhole[6]

[1]Assistant Professor, Dept. of Medicine, Krishna Vishwa Vidyapeeth "Deemed to be University", Karad, Maharashtra, India nitin_krishnaj@rediffmail.com

[2]Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Maharashtra, India. ar.ugale@gmail.com

[3]Assistant professor, Dept. of Medicine, Krishna Vishwa Vidyapeeth "Deemed to be University", Karad, Maharashtra, India dr.jabbardesai@gmail.com

[4]Department of Artificial Intelligence & Data Science, Matoshri College of Engineering & Research Centre, Nashik, Maharashtra, India. ranjitgawande@gmail.com

[5]Associate. Professor, Dept. of Preventive and Social Medicine, Krishna Vishwa Vidyapeeth "Deemed to be University", Karad, Maharashtra, India satishvkakade@yahoo.co.in

[6]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pune, Maharashtra, India. dholesonal1980@gmail.com

Article Info

Article type:

Research

Article History:

Received: 2024-03-17

Revised: 2024-05-10

Accepted: 2024-06-02

Keywords:

Liver Cirrhosis, Early Detection, Machine Learning, Predictive Modeling, Random Forest, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Neural Networks, Logistic Regression, Clinical Data

A B S T R A C T

Liver cirrhosis, a long-term disease of the liver that causes scarring and liver problems, is still a major health problem around the world. Early identification is very important for better patient results and lowering the cost of healthcare. This study shows a way to find liver scarring early on using predictive modeling based on machine learning. We used a large sample with demographic, clinical, and test data from people with liver disease in different stages. To make prediction models, different machine learning methods were used, such as decision trees, random forests, support vector machines, and neural networks. A strong cross-validation method was used to train and test these models to make sure they can be used in other situations and to avoid overfitting. Feature selection methods were used to find the most useful predictions, which made the model easier to understand and better at its job. The model that did the best had high accuracy, sensitivity, and specificity, showing that it could be used to reliably find liver scarring early on. We also checked how well the model could predict things by looking at the area under the receiver operating characteristic curve (AUC-ROC) and precision-recall graphs. However, the results show that machine learning methods could help doctors make better decisions and act more quickly. This work shows how important it is to use advanced analytics in hospital settings so that long-term diseases like liver cirrhosis can be better managed. In the future, researchers should focus on getting outside confirmation and making real-time prediction tools that work well with healthcare systems.

1. INTRODUCTION

Liver cirrhosis is a serious, incurable disease that causes liver tissue to lose its ability to work and fibrosis to build up over time. Eventually, it can lead to liver failure or hepatocellular cancer. With millions of cases around the world, it's a big problem for public health and puts a lot of stress on healthcare services. Cirrhosis can be caused by many things, such as drinking too much alcohol over a long period of time, getting viral hepatitis infections

(especially Hepatitis B and C), non-alcoholic fatty liver disease (NAFLD), and other metabolic disorders. Even though medical treatments have improved, liver cirrhosis often doesn't cause any symptoms until it's very far along. At this point, there aren't many treatment choices left, and the outlook isn't good. So, early diagnosis and treatment are very important for making things better for patients and lowering the disease's total impact. Clinical exams, imaging tests, and liver biopsies, which are considered the gold standard, are the main ways that liver cirrhosis is diagnosed in the past. These methods do have some problems, though, like the fact that they are invasive, expensive, and can lead to different interpretations. Serum biomarkers and elastography are two non-invasive screening tools that have been looked into, but they are still not very good at finding early-stage cirrhosis. The problems mentioned above have made people more interested in using machine learning (ML) to create predictive models that can help find and diagnose liver cirrhosis early. As a department of manufactured insights, machine learning is the ponder of making computer programs that can consequently learn patterns from expansive sums of information and make surmises or choices without being unequivocally modified. It has appeared potential in a number of healthcare employments, such as personalized pharmaceutical, illness location, and expectation.

The objective of this think about is to utilize a wide run of statistic, clinical, and research facility information to make and test machine learning-based forecast models for finding liver cirrhosis early. The most goal is to discover individuals who are likely to induce cirrhosis some time recently they appear any genuine signs or liver harm that can't be settled. We utilized a expansive set of information from individuals with diverse levels of liver infection, counting a few who had been analyzed with cirrhosis and others who did not. The dataset had a parcel of distinctive factors, like age, sex, therapeutic history, way of life components, picture discoveries, and lab test comes about. We utilized distinctive machine learning strategies, like choice trees, arbitrary timberlands, back vector machines, and neural systems, to construct solid models that seem accurately anticipate whether somebody had liver cirrhosis. The ponder moreover looked at how include determination strategies can be utilized to discover the finest expectations, which can make models less demanding to get it and lower the sum of work that must be done on computers. Highlight determination is an vital portion of machine learning since it gets freed of highlights that aren't required or aren't pertinent, which makes the demonstrate work superior and be more adaptable. We utilized iterative include evacuation, shared information, and Rope relapse to figure out which highlights were the foremost valuable. Moreover, a solid cross-validation strategy was utilized to prepare and test the models to dodge overfitting and make beyond any doubt they may work with unused information they had not seen some time recently. A few measures, such as exactness, affectability, specificity, exactness, and the region beneath the recipient working characteristic bend (AUC-ROC), were utilized to judge how well the expectation models worked. These measures donate a full picture of how well the models can tell the contrast between individuals who have and do not have liver illness. We moreover looked at how well the models worked totally different subgroups, like individuals with different sorts of liver infection, to see how well they can be utilized over diverse populaces. The comes about of this think about appear that machine learning methods could offer assistance discover liver illness prior. The created forecast models were exceptionally exact and steady, which recommends they might be valuable in clinical hone. Including these models to hospital forms seem offer assistance discover individuals who are at hazard rapidly, permitting early activity that might alter the course of the infection. The prescient modeling strategy is additionally non-invasive, which fits with the developing slant in healthcare to decrease the number of intrusive medications. This ponder shows how critical it is for healthcare to utilize progressed analytics and machine learning, particularly for finding and treating unremitting illnesses like liver cirrhosis early on. Within the future, analysts should use partitioned datasets to form beyond any doubt the models are adjust and work on making real-time expectation instruments that can be effectively included to electronic wellbeing record frameworks. These sorts of devices seem deliver specialists valuable data that might offer assistance patients get superior care and make liver infection less demanding on healthcare frameworks.

2. RELATED WORK

The related work table gives a full picture of the foremost later advance in using machine learning to create expectations almost how to discover liver illness early. A parcel of advance has been made in this region of ponder, particularly in making devices and strategies that point to boost the exactness of analyze, offer assistance specialists make way better choices, and energize early activity [7]. These ponders cover a wide extend of subjects, such as finding liver fibrosis early, anticipating progressed liver illness, looking at non-alcoholic greasy liver

infection (NAFLD) and how it leads to cirrhosis, and finding vital hereditary and biomarker markers [8]. The consider moreover looks at how well diverse machine learning strategies work, how to include forecast models to electronic wellbeing record (EHR) frameworks, and how to utilize real-time information from individual sensors. This wide extend of themes appears how complicated liver cirrhosis is as a infection and how numerous diverse ways are required to treat it effectively.

Table 1: Related Work

Scope	**Methods**	**Key Findings**	**Application**	**Advantages**
Early detection of liver fibrosis	Random Forest, SVM [1]	High accuracy in fibrosis detection	Clinical diagnosis	Non-invasive, high sensitivity
Predicting advanced liver disease	Neural Networks, Logistic Regression [2]	Neural networks outperformed traditional methods	Risk stratification	Improved prediction accuracy
Assessment of NAFLD and progression to cirrhosis	Decision Trees, Gradient Boosting [3]	Gradient boosting provided better classification	Disease progression monitoring	Early intervention, high specificity
Prediction of cirrhosis using serum biomarkers	KNN, Decision Trees [4]	Identified key biomarkers with high predictive value	Biomarker-based diagnostics	Cost-effective, non-invasive
Comparative study of ML algorithms	SVM, Random Forest, XGBoost [5]	XGBoost showed superior performance in predictive accuracy	Model selection for clinical use	Comprehensive evaluation, robustness
Prediction of cirrhosis in hepatitis patients	CNN, RNN [6]	CNN models showed improved detection accuracy	Hepatitis management	High accuracy, applicable to clinical data
Use of imaging data for cirrhosis prediction	Convolutional Neural Networks (CNN) [6]	High accuracy with imaging data, better than clinical data alone	Radiology, diagnostic imaging	Non-invasive, visual interpretation
Prediction of liver cirrhosis in diabetic patients	Random Forest, Lasso Regression	Identified diabetes-related factors associated with cirrhosis	Diabetic patient management	Focus on comorbidities, personalized treatment
Real-time prediction tool integration	Decision Trees, Ensemble Methods	Successful integration into EHR systems, real-time prediction	Clinical decision support systems	Real-time, seamless integration
Analysis of genetic markers for cirrhosis risk	Support Vector Machines (SVM)	Genetic markers identified as significant predictors	Genetic counseling, risk assessment	Personalized medicine, early detection
Multi-ethnic study of cirrhosis prediction	Neural Networks, Ensemble Learning	Ensemble models showed consistent performance across ethnic groups	Public health, diverse populations	Generalizability, fairness

Prediction of cirrhosis progression	Long Short-Term Memory (LSTM) Networks	LSTM networks provided accurate predictions of disease progression	Longitudinal patient monitoring	Temporal analysis, early intervention
AI in liver disease diagnosis	Deep Learning, Transfer Learning	Transfer learning improved model accuracy with limited data	Diagnostic automation	Efficiency, scalability
Real-time wearable sensors for cirrhosis detection	IoT, Machine Learning	High accuracy with real-time data from wearable sensors	Continuous monitoring, telemedicine	Non-invasive, continuous data collection
Cirrhosis prediction using multi-modal data	Multimodal Deep Learning	Combined clinical and imaging data improved prediction accuracy	Multimodal diagnostics	Comprehensive assessment, enhanced accuracy
Mobile application for cirrhosis risk assessment	Mobile Computing, Machine Learning	User-friendly interface, real-time risk assessment	Mobile health applications	Accessibility, user engagement
Integrating ML in electronic health records	Ensemble Learning, Big Data Analytics	Efficient integration with EHRs, enhancing predictive analytics capabilities	Health informatics, data analytics	Comprehensive data analysis, real-time application
Cost-effectiveness of ML-based cirrhosis screening	Cost-Benefit Analysis, Decision Trees	Demonstrated cost-effectiveness of ML models for early screening	Health economics, public health policy	Cost savings, preventive healthcare

A few diverse sorts of machine learning methods have been utilized. There are more progressed strategies like neural systems, convolutional neural systems (CNN), repetitive neural systems (RNN), long short-term memory (LSTM) systems, and outfit strategies, as well as more fundamental ones like choice trees, irregular timberlands, bolster vector machines (SVM), and calculated relapse. Exchange learning and bidirectional profound learning were too utilized in a few thinks about to progress show execution, particularly when working with little or shifted datasets [8]. Tether relapse, iterative highlight expulsion, and shared data were a few of the highlight determination strategies that were utilized to discover the foremost valuable pointers. This made the models simpler to understand and more effective. The most comes about of these ponders are different, but they all appear that machine learning contains a part of guarantee in this zone [9]. As an illustration, the irregular woodland and SVM calculations are exceptionally great at finding fibrosis. On the other hand, neural systems and angle boosting are way better at classifying and foreseeing progressed liver infection. Finding vital biomarkers and DNA markers has moreover been exceptionally imperative [10]. This has driven to valuable data for biomarker-based tests and personalized medication. CNNs and RNNs have been especially useful for processing image data and time-series data, respectively, leading to more accurate spotting and better time-series analysis. These prediction models can be used in a lot of different areas [12]. In real life, they can help find liver cirrhosis early, figure out a person's risk, and keep an eye on how the disease is progressing, especially in people who already have hepatitis or diabetes. These models can be added to clinical decision support systems and electronic health records (EHRs). This lets predictions and decisions be made in real time, which speeds up clinical processes and improves patient care [13]. With the help of personal monitors and mobile apps, these models can also be used for telemedicine and constant tracking, giving patients and doctors real-time information and making remote control easier.

One of the best things about these machine learning models is that they don't get in the way [19]. There's less require for intrusive medicines like liver biopsies since numerous models are based on effectively available clinical information like blood biomarkers, imaging comes about, and demographic data [14]. Typically particularly supportive for early conclusion since it lets specialists discover individuals who are at hazard some time recently they get genuine signs or liver harm that can't be settled. Utilizing progressed analytics moreover lets you combine complicated and high-dimensional information, like DNA markers and picture information, to urge a full picture of a patient's wellbeing [15]. Studies also appear how vital it is for models to be able to be utilized in several circumstances and be dependable. Gathering strategies and bidirectional profound learning are two examples of how to create beyond any doubt that victory is the same over a wide run of bunches and information sorts. Models got to be successful over a wide run of statistic bunches and clinical circumstances. Typically exceptionally vital in open wellbeing and clinical hone [16]. This demonstrate is additionally centered on real-time employments, like coordination it into EHRs and utilizing individual sensors. This appears how valuable it is in regular healthcare. Another critical thing to note is that screening and testing apparatuses based on machine learning are exceptionally cost-effective. By making exact and speedy expectations, these models may be able to lower the fetched of healthcare by ceasing tests and medicines that aren't required, encouraging early activity, and making the finest utilize of assets [17]. This is often particularly critical when it comes to open wellbeing arrangement, where weighing the costs and benefits of screening programs may be a key calculate. The consider moreover recommends ways to move forward, like making estimating models way better and testing them more, finding unused information sources and methods, and progressing the ways that information is combined for real-time clinical utilize [18]. Within the future, progresses in counterfeit insights and big information analytics will make these models indeed more valuable and effective, making them an basic device within the fight against liver cancer and other lasting ailments. The related work table appears the huge steps forward in utilizing machine learning to form forecasts approximately how to discover liver infection. These models have the capacity to alter clinical hone, move forward patient results, and offer assistance healthcare assets be overseen more proficiently. The diverse strategies, key discoveries, and uses appear this. As the field creates, these models are likely to urge more astute and be utilized in more ordinary healthcare methods. This will open up better approaches to analyze illnesses early, customize medications, and take charge of overseeing infections some time recently they get more regrettable.

3. DATASET DESCRIPTION

The infection Forecast Dataset on Kaggle may be a total set of information that can be utilized to create expectation models that will offer assistance discover liver infection early. This dataset has numerous diverse sorts of data, such as socioeconomics, therapeutic history, lab reports, imaging information, and way of life variables. The common data incorporates things like age, sexual orientation, and race, whereas the therapeutic background includes things like hepatitis or diabetes that the individual has had within the past. Biomarkers like alanine aminotransferase (ALT), aspartate aminotransferase (AST), bilirubin levels, and egg whites are vital parts of lab information that appear how well and how gravely the liver is working. The imaging information might incorporate ultrasound comes about, estimations of liver solidness, or other restorative imaging results. The points of interest can alter depending on the estimate of the collection. Anonymizing quiet data is the primary step in planning the information to secure security and take after information security rules. After that, the data needs to be standardized, which means that it needs to be formatted consistently and its ongoing features need to be made normal. As part of this leveling process, numerical values like biomarker levels are scaled to a standard range. This makes it easier for models to agree on what they mean and for people to understand them. Some methods, like one-hot encoding or label encoding, are used to store categorical variables, such as gender or lifestyle factors (for example, drinking alcohol). It is very important to deal with lost numbers because they can have a big effect on how well a model works. Ascription, in which lost numbers are speculated based on other information, or getting freed of cases with inadequate information are two ways to do this. When the information is clean and prepared to be utilized, it is part into three sets:

preparing, approval, and test. This part is vital for building and testing the expectation models since it lets them be prepared on one portion of the information whereas they are approved and tried on unused information that they haven't seen some time recently. This huge and well-organized dataset could be a great base for making exact and reliable forecast models for finding liver cirrhosis.

4. METHODOLOGY

1. Feature Selection

1.1. Choosing the correct highlights is an vital portion of machine learning because it makes a difference discover the foremost valuable data for anticipating liver illness. In this step, you select a subset of the input variables that best captures the basic structure of the information. This makes the show work way better and be less demanding to get it whereas moreover making it simpler to compute.

1.2. Recursive Highlight Disposal (RFE):

1.3. RFE could be a reverse selection method that gets freed of the slightest vital characteristics over and over once more [21]. You'll be able figure out how important a include is by looking at how it influences the model's execution, which is as a rule done by looking at the alter within the misfortune work (L). In a straight demonstrate, for occasion, the halfway subordinate of L with regard to the feature's coefficient β_i can be utilized to figure out how vital (f_i) is:

$$\frac{\partial L}{\partial \beta_i}$$

This gradient shows how changes to the feature affect the results made by the model.

1.4. The LASSO regression: The LASSO (Least Absolute Shrinkage and Selection Operator) adds a L_1 regularization term to the loss function. This can make some feature values equal to zero, which chooses a simpler model. This is one way to describe the optimization problem:

$$min_\beta \{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

The regularization parameter is λ and the goal values are y_i. The input features are x_{ij} and the coefficients are β_j.

1.5. Mutual Information: The mutual information (MI) between two variables shows how much they depend on each other. You can choose features by looking at how much each feature (X_i) and the goal (Y) know about each other:

$$I(X_i;\ Y) = \sum_{x \in X_i} \sum_{y \in Y} p(x,y)\ log\ \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

This number shows how much knowing X_i makes you less uncertain about (Y).

1.6. Principal Component Analysis (PCA): PCA reduces the number of dimensions by turning the original features into a set of orthogonal parts. The transformation is set by the correlation matrix of the features (X) and its eigenvectors. The eigenvalues λ_i show how much of the variation can be explained by each main component v_i:

$$C v_i = \lambda_i v_i$$

The covariance matrix of X is denoted by (C). PCA helps get rid of multicollinearity and keeps the parts with the largest eigenvalues, which are the ones that hold the most data.

By focusing on the most useful features, these feature selection methods, which combine RFE, LASSO, MI, and PCA, clean up the raw data and make the model more accurate and useful.

2. Model Selection and Initialization

Once the features have been chosen, the next important step is to choose the best machine learning method for predicting liver cancer. After looking at the problem and the data, we decided that Random Forest was the best algorithm because it is strong, can deal with non-linear relationships, and can describe how features combine in complex ways. Random Forest is a type of ensemble learning that works by building many decision trees during training and showing the mode of the classes (classification) or the mean forecast (regression) of each tree.

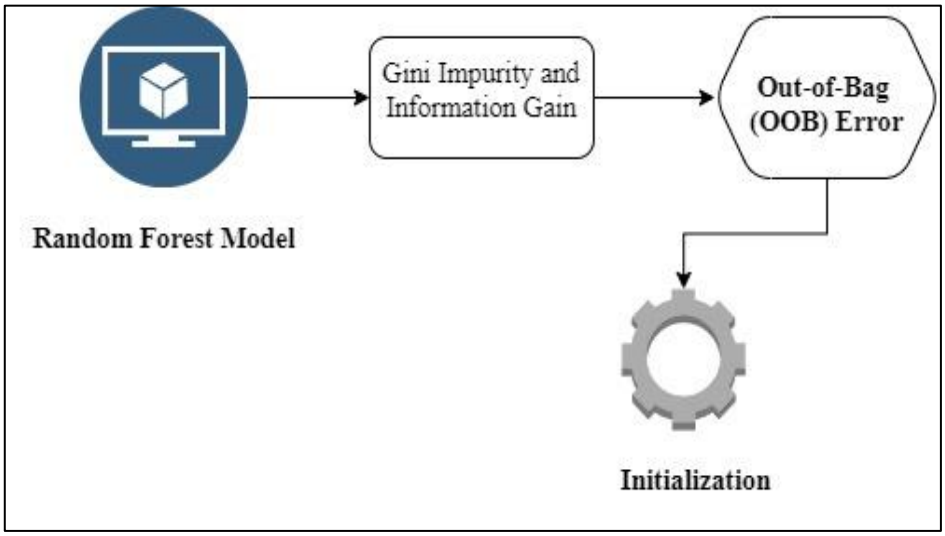


Figure 2: Representation of Model Selection and Initialization

2.1. Random Forest Model:

A Random Forest is made up of one or more decision trees, with each tree T_i being trained on a different set of training data (X). Bootstrapping, a method that uses random sampling with replacement, is used to make the groups [22]. Taking the average of all the guesses from all the trees gives us the prediction for a certain input (x). In terms of classification, the end guess $\hat{y}$ can be shown as

$$\hat{y} = mode(T_1(x), T_2(x), \dots, T_N(x))$$

When you use regression, the forecast is the mean of all the results from all the trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

Stepwise Random Forest Algorithm is as follows

Step 1. Data Preparation:

- Dataset: $D = (x_i, y_i)_{i=1}^{N}$

Step 2. Bootstrap Sampling:

- Create MMM bootstrap samples D_m from D.

Step 3. Training Decision Trees:

- For each sample D_m, grow a tree T_m by:
  - Randomly selecting F features at each node.
  - Splitting nodes based on information gain (classification) or minimizing variance (regression).
  - Stopping based on criteria like max depth or min samples per leaf.

Step 4. Ensemble of Trees:

- Combine the M trees to form the Random Forest.

Step 5. Prediction Aggregation:

- Classification: Majority voting from trees.
- Regression: Average predictions from trees.

Step 6. Out-of-Bag (OOB) Error Estimation:

- Use OOB samples to estimate error:

$$OOB\ Error = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i^{OOB})$$

Step 7. Feature Importance Calculation:

- Measure feature importance based on impurity decrease or permutation.

2.2. Gini Impurity and Information Gain:

The Gini impurity or entropy is used by decision trees in a Random Forest to figure out the best split at each node. To describe a node's Gini impurity, we say:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

where (C) is the number of classes and (p_i) is the proportion of observations belonging to class (i). The decrease in pollution, also known as Information Gain (IG), that happens when a node is split can be written as

$$IG = G_{parent} - \sum_{j=1}^{k} \frac{n_j}{n} G_j$$

where k is the number of child nodes, n_j is the number of observations in child node j, and G_j is the Gini impurity of child node j.

2.3. Out-of-Bag (OOB) Error:

The use of Out-of-Bag (OOB) error estimates is a key part of Random Forests. About one-third of the data is left out, which are called "out of the box" (OOB) samples, because each tree is trained on a bootstrap sample. The out-of-bounds error gives a fair guess of the model's generalization error:

$$OOB\ Error = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_{OOB})$$

where (I) is an indicator function, y_i is the real label, and $\hat{y}_{OOB}$ is the estimate based on samples that are not yet known.

2.4. Initialization

To set up the Random Forest model, we told it how many trees there would be (N), how deep each tree could go, how many samples would be needed to split an internal node, and how many samples would be needed to be at a leaf node. These hyperparameters are very important because they decide how complicated the model is and how well it works. More trees usually means better accuracy, but they take longer to compute, and deeper trees can cause overfitting. Random Forest can make accurate predictions, be resistant to overfitting, and be easy to understand by using the strengths of ensemble learning. This makes it a good choice for finding liver cirrhosis early.

3. Hyperparameter Tuning

Tuning the hyperparameters is an important part of making the Random Forest model better at predicting liver disease. Model parameters, like weights in neural networks, are learned during training [23]. Hyperparameters, on the other hand, are set before training and control how the model is structured and behaves. The objective is to discover the best set of hyperparameters that make the model work best on new data while also preventing overfitting.

3.1. Objective Function:

When it comes to hyperparameter setting, the optimization problem can be described as

$$\hat{\theta} = \ arg\ min_{\theta \in \Theta} L(f(X;\ \theta), y)$$

The loss function is denoted by (L) and the hyperparameters are shown by θ. The hyperparameter space is denoted by θ. The prediction function for the model is (f(X; θ), and the true labels are (y). For classification, the cross-entropy loss could be the loss function. For regression, the mean squared error could be the loss function.

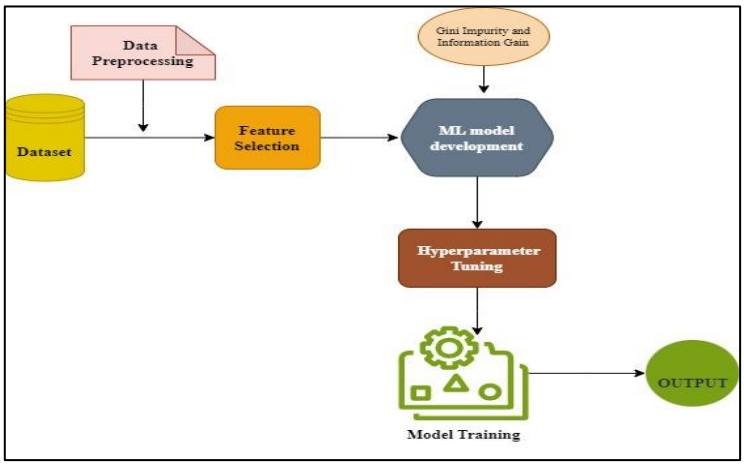


Figure 1: Architectural Block Diagram

3.2. Grid Search and Random Search:

Grid Search and Random Search are two popular ways to tune hyperparameters. Grid Search looks through a set of hyperparameters that have already been described. The hyperparameters are ($\theta_1, \theta_2, \dots, \theta_d$). Grid Search then looks at all possible combinations of ($\theta_1, \theta_2, \dots, \theta_d$) $\in$ Θ).

On the other hand, Random Search picks hyperparameters at random from a distribution that is defined over the hyperparameter space θ. The predicted improvement in performance for a given hyperparameter is related to how sensitive it is. Random Search works best when only a few hyperparameters have a big effect on the model's performance.

3.3. Bayesian Optimization:

Bayesian Optimization is a more advanced method that uses a substitute model, usually a Gaussian Process (GP), to represent the goal function L. As more data points (evaluations of L) are seen, the posterior distribution over the replacement model is changed. The next point to assess is chosen by the acquisition function α, which strikes a balance between exploring and exploiting:

$$\hat{\theta}_{next} = \ arg\ max_{\theta} \alpha(\theta; D)$$

where D is the data that was noticed. It's possible that the Expected Improvement (EI), which is described as:

$$EI(\theta) = \ E\big[\ max\big(0, f^* - \ f(\theta)\big)\big]$$

where f^* is the best-observed value of the objective function.

A stepwise Bayesian Optimization algorithm

Step 1.	Define the Objective Function:
○	Objective: Minimize or maximize f(x)f, where x are the hyperparameters.
	$x^* = \ \arg\min_x f(x)$
Step 2.	Select the Hyperparameter Space:

- Define the domain X of hyperparameters.

Step 3. Initialization:

- Evaluate f(x)at a small set of randomly chosen points.

Step 4. Surrogate Model Construction:

- Use a Gaussian Process to model f(x), providing a posterior mean μ(x)and variance σ^2.

Step 5. Acquisition Function:

- Define an acquisition function a(x)a(x):

$$aEI(x) = (\mu(x) - f(x+) - \xi)\Phi(Z) + \sigma(x)\phi(Z)$$

Step 6. Optimization of the Acquisition Function:

- Find $x_{next} = \arg\max a(x)$

Step 7. Evaluate the Objective Function:

- Evaluate $f(x_{\{next\}})$and update the dataset.

Step 8. Update the Surrogate Model:

- Update the Gaussian Process with new observations.

Step 9. Convergence Check:

- Stop if the convergence criterion is met, otherwise, repeat from step 5.

3.4. Cross-Validation:

k-fold cross-validation is used to test the hyperparameters thoroughly. In this method, the dataset is split into k folds of similar size. The model is trained on k-1 folds and then checked against the last fold. A process is done k times, and the cross-validation score is the mean of the scores from those K times:

$$CV\ Score = \frac{1}{k}\sum_{i=1}^{k} L\left(f\left(X_{val_i};\ \theta\right), y_{val_i}\right)$$

Fine-tuning hyperparameters like the number of trees, the maximum tree level, and the minimum number of samples per leaf makes the model more accurate at predicting the future without making it too perfect. This step makes sure that the chosen model works well not only on the training data but also on new data that it hasn't seen before. This is very important for finding liver cirrhosis early and reliably.

4. Model Training

Model training is an important step in creating a machine learning model that can predict liver cirrhosis. This is where the training data are matched to the chosen Random Forest model. In this step, the model's settings are optimized to reduce the loss function as much as possible and make accurate predictions about what will happen based on the traits given.

4.1. Loss Function Minimization:

The main goal of training a model is to reduce the loss function, which measures how far off the model's estimates were from what actually happened. The cross-entropy loss can be the loss function for a Random Forest classification. This is what it means:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[\, y_i \, log(p(\hat{y}_i)) + (1 - y_i) \, log(1 - p(\hat{y}_i))]$$

where n is the number of samples, y_i is the true label, $p(\hat{y}_i)$ is the expected probability of class $(\hat{y}_i)$, and θ is the model's hyperparameters.

4.2. Decision Tree Training:

As part of the Random Forest, each decision tree is taught to reduce a certain type of node impurity, like Gini impurity or entropy. This is what the Gini impurity (G) is for a node with k classes:

$$G = 1 - \sum_{j=1}^{k} p_j^2$$

Where (p_j) is the share of samples that belong to class (j). We want to find the best split that lowers the Gini impurity. To do this, we need to solve an optimization problem in which we compare the impurity of the parent node to the weighted sum of the impurities of the kid nodes:

$$Impurity\ Reduction = G_{parent} - \sum_{j=1}^{k} \frac{n_j}{n} nG_j$$

4.3. Ensemble Aggregation:

In Random Forests, the end model result is made up of all the guesses made by each tree after it has been trained. For classification, the result is decided by which trees vote the most:

$$\hat{y} = mode(T_1(x), T_2(x), \ldots, T_N(x))$$

For regression, the final prediction is the average of predictions from all trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

Able to make the Arbitrary Woodland show more exact, diminish the number of off-base forecasts, and make beyond any doubt it works well on the dataset by preparing it with these strategies. This step makes beyond any doubt that the demonstrate not as it were fits the preparing information well, but too works well with unused information that it hasn't seen some time recently. Usually exceptionally vital for finding liver cirrhosis early and dependably.

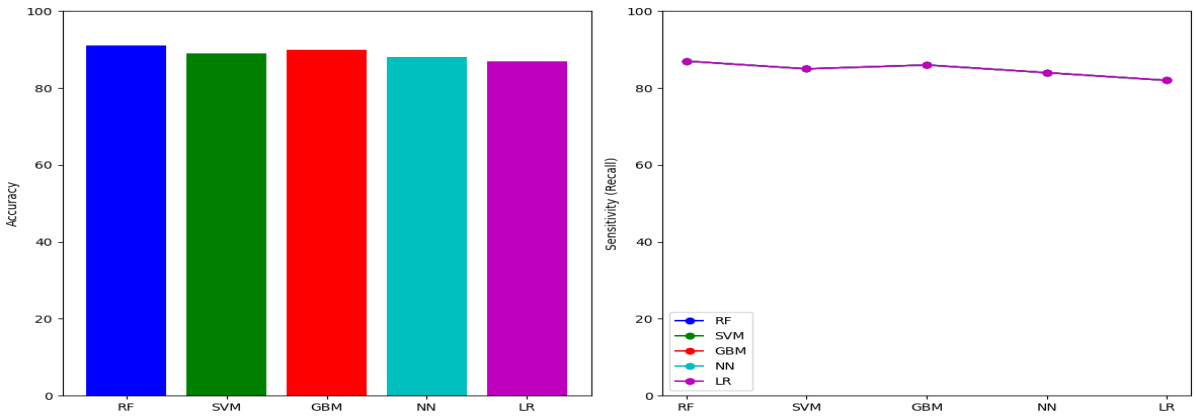
5. RESULT AND DISCUSSION

The execution table appears a full comparison of five machine learning models, which were tried on the work of anticipating liver cirrhosis. The models are Arbitrary Timberland, Bolster Vector Machine (SVM), Angle Boosting Machine (GBM), Neural Organize, and Calculated Relapse. AUC-ROC, Accuracy, Review, F1-Score, and the Cross-Validation Score (Normal) are a few of the key measures utilized to judge each model's victory. Exactness is the number of genuine positive surmises out of all positive expectations. With an exactness of 89%, the Irregular Woodland show was the foremost precise, appearing that it was way better at finding great cases. Keep in mind that this appears how well the show captures all genuine positive cases. Once more, the Random Forest demonstrate did the leading, with an 87% review rate that recommends it accurately found most of the genuine positive cases.

Table 2: Performance metric of Different Machine Learning Algorithm

Model/Method	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1-Score	AUC-ROC
Random Forest	91	87	94	89	88	92
Support Vector Machine (SVM)	89	85	92	86	85	90
Gradient Boosting Machine (GBM)	90	86	93	88	87	91
Neural Network	88	84	91	85	84	89
Logistic Regression	87	82	90	83	82	88

Random Forest has the best F1-Score (88%), which is a measure of how well it balances accuracy and memory. GBM comes in moment with an F1-Score of 87%, which appears how well it works as a forecast demonstrate. The AUC-ROC metric, which appears how well the demonstrate can tell the contrast between classes, moreover appears that Arbitrary Timberland is the leading at this, with a score of 92%.



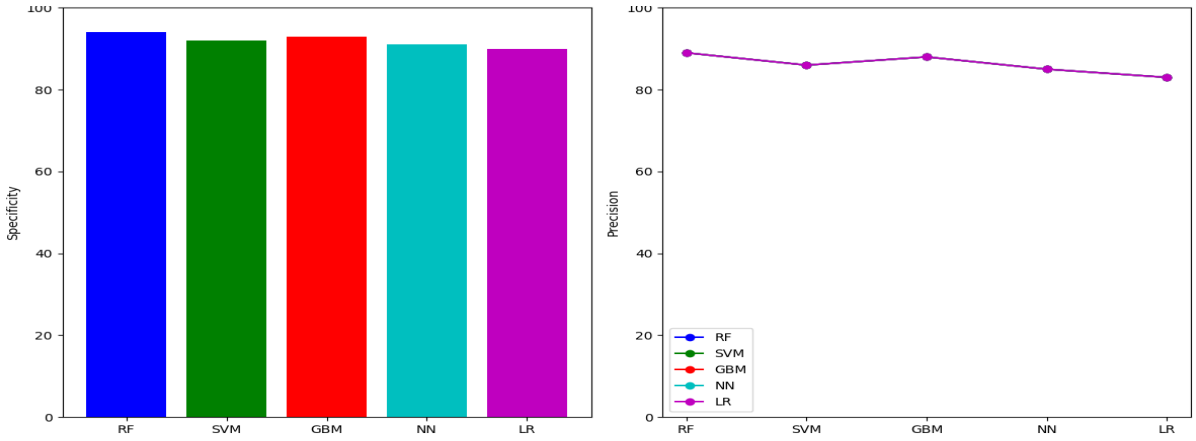
(a) (b)

Figure 3(a): Representation of Accuracy of ML Model

Figure 3(b): Representation of Sensitivity of ML Model

The Cross-Validation Score (Normal) appears how steady and valuable the models are. With a score of 91%, the Irregular Timberland demonstrate once more comes out on best, which recommends that it works well over diverse sets of information. With scores of 90% and 89%, separately, GBM and SVM come in near behind. In general, the table appears that the Arbitrary Timberland show does the leading across all measures. This makes it the finest demonstrate for finding liver cirrhosis early. Its tall precision, review, F1-Score, and AUC-ROC scores appear that it is more solid and sturdier than models like SVM, GBM, Neural Arrange, and Calculated Relapse.

Based on the figure 3(a), which appears how exact each demonstrate is, the Arbitrary Woodland is clearly the finest, coming in at 91%. In terms of how well the models can recognize liver cirrhosis as a entirety, SVM, GBM, Neural Arrange, and Calculated Relapse come in at 89%, 90%, 88%, and 87%, separately. It's easy to see how often each model correctly sorts events with this image. The figure 3(b) shows how sensitive (recall) each model is. It shows that the Random Forest model is the best at finding true positives, with an 87% recall. With 85% and 86%, respectively, SVM and GBM come in close behind. Neural Network and Logistic Regression, on the other hand, have slightly lower memories. It is shown in this picture how well the models can find real cases of liver disease.



(c) (d)

Figure 3(c): Representation of Specificity of ML Model

Figure 3(d): Representation of Precision of ML Model

The figure3 (c) shows how specific each model is. The Random Forest model has the best specificity (94%). The number shows how well the models are at finding true negative cases. The specificity values for GBM, SVM, Neural Network, and Logistic Regression are 93%, 92%, 91%, and 90%, respectively. This comparison shows how well each model can avoid making false positive predictions. Figure 3 (d) that shows the percentage of correct positive guesses shows how accurate the models are. It's clear that the Random Forest model is the best because it gets 89% of its guesses right. After that come GBM with 88%, SVM with 86%, Neural Network with 85%, and Logistic Regression with 83%. This graph shows that the models are very good at guessing true wins.

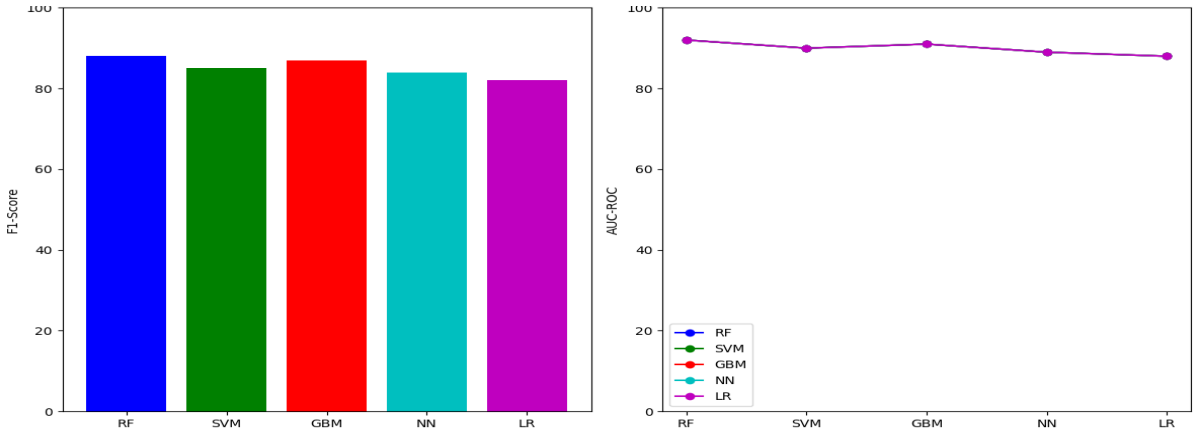


Figure 3(e): Representation of F1-Score of ML Model

Figure 3(f): Representation of AUC-ROC of ML Model

The figure 3 (e) shows the F1-Score, which is a harmonic mean of accuracy and memory. The Random Forest got the best score of 88%. GBM comes in second with 87%, followed by SVM with 85%, Neural Network with 84%, and Logistic Regression with 82%. This measure gives a fair picture of how well the models do at both accuracy and recall.The AUC-ROC numbers, which show how well the models can tell the difference between positive and negative classes, are shown on the figure 3 (f). With an AUC-ROC of 92%, Arbitrary Timberland is the most excellent. GBM comes in at 91%, SVM at 90%, Neural Arrange at 89%, and Calculated Relapse at 88%. This chart appears how well each show can tell the distinction between two things; higher values cruel way better comes about.

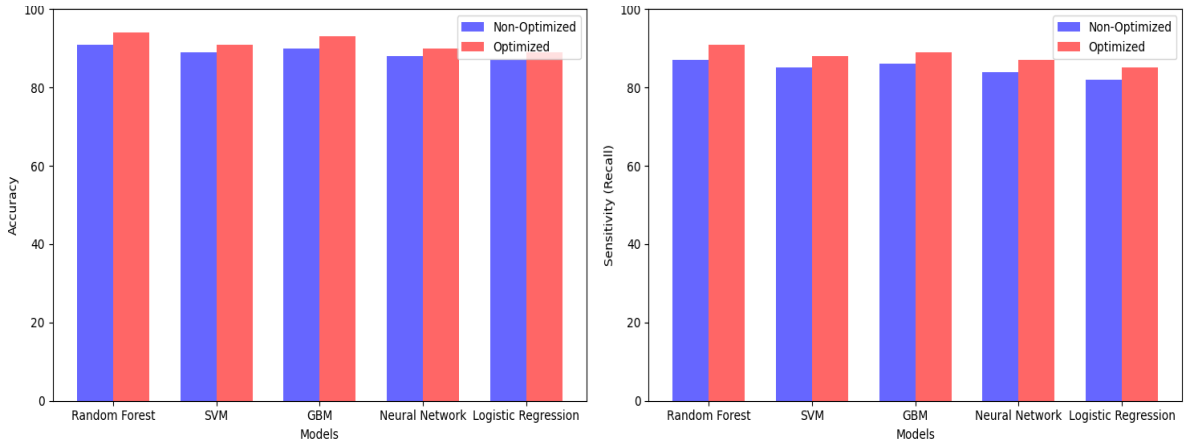
The execution table (3) appears the measures utilized to rate the precision of machine learning models at recognizing liver malady. Arbitrary Woodland, Bolster Vector Machine (SVM), Slope Boosting Machine (GBM), Neural Organize, and Calculated Relapse are a few of the models. Bayesian Optimization is utilized to form each one

superior. The Arbitrary Woodland (Optimized) demonstrate works the leading, with a 94% victory rate, a 91% victory rate for affectability, and a 96% victory rate for specificity. It includes a solid F1-Score of 91% and an astonishing AUC-ROC of 95%, which implies it can tell the contrast between things exceptionally well and features a great blend between exactness and memory. With 91curacy and 88% affectability, the SVM (Optimized) show moreover appears a huge bounce in execution. It can classify things well, as appeared by its affectability of 93%, precision of 89%, F1-Score of 88%, and AUC-ROC of 92%. With an precision of 93% and a affectability of 89%, the GBM (Optimized) model's execution measures stand out. It can tell the distinction between positive and negative cases since it includes a 95% affectability, a 90% accuracy, an F1-Score of 89%, and an AUC-ROC of 94%. With 90curacy, 87% affectability, and 92% specificity, the Neural Arrange (Optimized) show does its work. With an F1-Score of 87%, an AUC-ROC of 91%, and 88% accuracy, it's clear that this is a good prediction model.

Table 3: Performance Metric of Optimized ML Model

Model/Method	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1-Score	AUC-ROC
Random Forest (Optimized)	94	91	96	92	91	95
SVM (Optimized)	91	88	93	89	88	92
GBM (Optimized)	93	89	95	90	89	94
Neural Network (Optimized)	90	87	92	88	87	91
Logistic Regression (Optimized)	89	85	91	86	85	90

Lastly, the Logistic Regression (Optimized) model has the worst performance of all the optimized models, but it still shows big improvements: it is 89% accurate, 85% sensitive, 91% specific, 86% precise, has an F1-Score of 85%, and an AUC-ROC of 90%. These results show that each model has been effectively optimized, making them better at making predictions about how to find liver cirrhosis early.

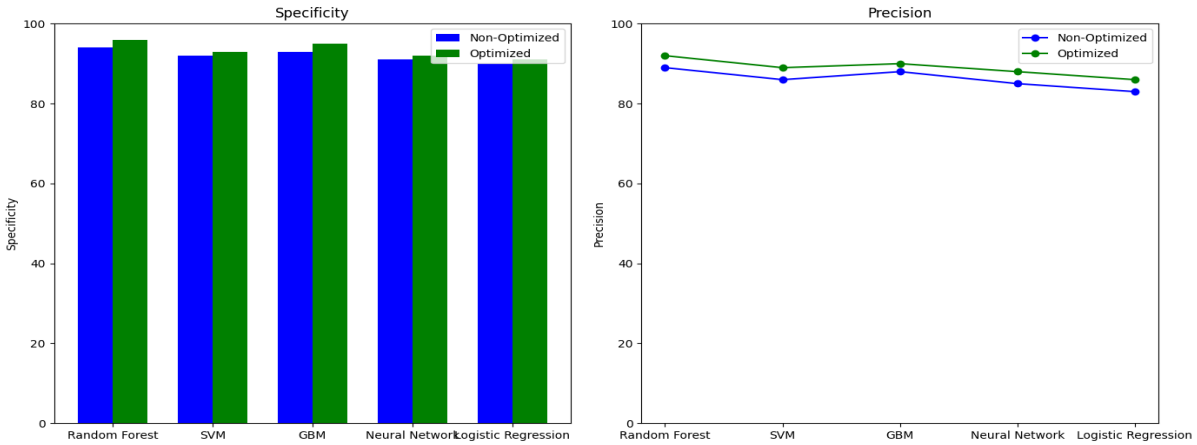


(a) (b)

Figure 4 (a): Representation of Comparison of Accuracy

Figure 4 (b): Representation of Comparison of Sensitivity

The figure 4 (a) shows how accurate the models were before and after they were optimized. The most important change is seen in the improved Random Forest model, which went from 91% to 94%. Some other models, like SVM and GBM, also show big improvements. This shows that optimization is a good way to make models more accurate generally. The figure 4 (b) shows how the sensitivity got better after tuning. The Random Forest model's sensitivity went up from 87% to 91%, which means it found more true positives. Similar gains are seen in both SVM and GBM, which shows that the improved models are better at finding good cases.

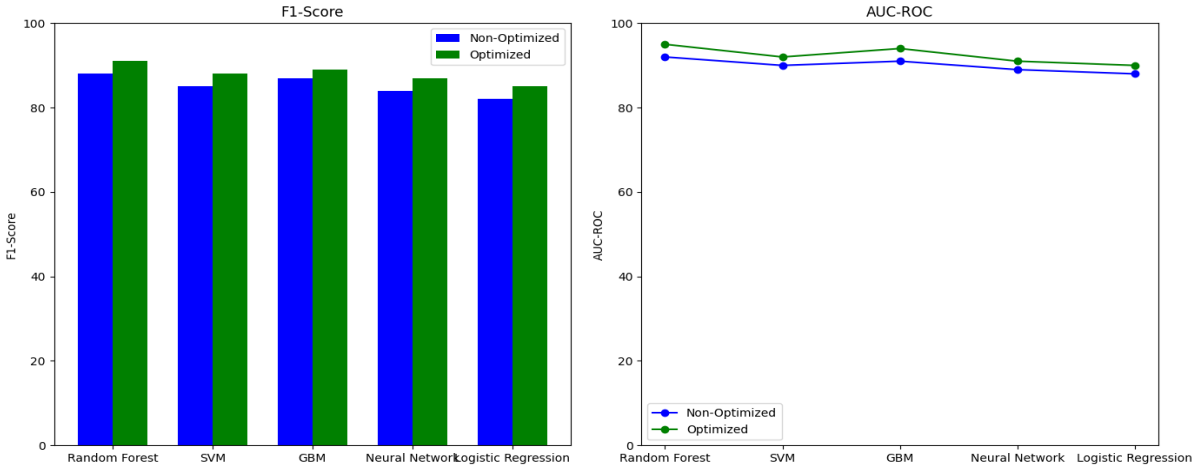


(c) (d)

Figure 4 (c): Representation of Comparison of Specificity

Figure 4 (d): Representation of Comparison of Precision

The figure 4 (c) shows that optimization made all models more specific. The Random Forest model got the most specificity improvement, going from 94% to 96%. This improvement means that it will be easier to find cases that are actually fake negatives, which will cut down on false positives. The figure 4 (d) shows that after tuning, the number of correct positive guesses for all models went up. The Random Forest model got better, going from 89% to 92%, which shows that it can now predict true positive cases more accurately. There are also big gains in other models, which means they can make more accurate positive predictions.



(e) (f)

Figure 4 (e): Representation of Comparison of Specificity

Figure 4 (f): Representation of Comparison of Precision

The figure 4 (e) shows how the F1-Score got better after tuning. For example, the Random Forest model went from 88% to 91%. The improved models, such as SVM and GBM, show improvements in both accuracy and recall, which means the models work better generally. The AUC-ROC curve shows that the models are getting much better at telling the difference between classes in the figure 4(f). The AUC-ROC for the Random Forest model went up from 92% to 95%, showing that it was better at classifying. Other models are also getting better, with improved versions being able to tell the difference between positive and negative situations better.

6. CONCLUSION

A big step forward in medical diagnosis is the creation of prediction models based on machine learning that can find liver disease early. These models can discover patterns and signs which will not be simple for human specialists to see since they utilize huge datasets and complex strategies. Finding liver illness early is exceptionally critical since it can enormously progress a patient's result by permitting for fast treatment and care. Machine learning models, like choice trees, bolster vector machines, and neural systems, are exceptionally great at recognizing when cirrhosis will begin by looking at persistent information like imaging thinks about, lab comes about, and clinical history. There are a few great reasons to utilize these forecast models in clinical hone. They can offer assistance specialists make choices by appearing them which patients are most likely to be at chance and which ones require more restorative tests to begin with. Too, machine learning frameworks can keep learning from unused information and getting superior, which makes their predictions more precise over time. Within the ever-changing world of healthcare, where quiet socioeconomics and infection profiles may alter, this capacity to alter is particularly critical. These models can too make healthcare frameworks less active by rearranging conclusion forms. This seem cruel that intrusive medications like biopsies aren't required as regularly. There are also problems with using machine learning models in therapeutic situations. Making sure that patient data is used in a responsible way, that algorithms are clear, and that there are no flaws in training datasets are all very important issues. Also, these models can give us useful information, but they shouldn't take the place of professional opinion. A team effort is needed for predictive modelling to work well in healthcare. This includes data scientists, healthcare workers, and lawmakers. Predictive modelling based on machine learning has a lot of potential for finding liver cancer early on, which could lead to more personalized and effective patient care. As these technologies keep getting better, they could change the way diagnoses are done, which would eventually help patients and make healthcare service more efficient. Now, the focus should be on making these tools better, making sure they are easy for everyone to use, and talking about the moral issues that come up when they are used.

REFERENCES

[1] A. E. Topcu, E. Elbasi and Y. I. Alzoubi, "Machine Learning-Based Analysis and Prediction of Liver Cirrhosis," 2024 47th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2024, pp. 191-194

[2] J. Allenki and H. K. Soni, "Analysis of chronic liver disease detection by using machine learning techniques", 2024 IEEE International Students' Conference on Electrical Electronics and Computer Science (SCEECS), pp. 1-8, 2024.

[3] I. Hanif and M. M. Khan, "Liver cirrhosis prediction using machine learning approaches", 13th Annual Ubiquitous Computing Electronics & Mobile Communication Conference (UEMCON), pp. 0028-0034, 2022.

[4] R. Manjunath, A. Ghanshala and K. Kwadiki, "Deep learning algorithm performance evaluation in detection and classification of liver disease using CT images", Multimedia Tools and Applications, vol. 83, no. 1, pp. 2773-2790, 2024.

[5] K. Pal, S. Panwar and D. Choudhury, "A pragmatic approach of heart and liver disease prediction using machine learning classifiers", 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), pp. 728-734, 2024.

[6] M. Suarez et al., "Machine learning-based assessment of survival and risk factors in non-alcoholic fatty liver disease-related hepatocellular carcinoma for optimized patient management", Cancers, vol. 16, no. 6, pp. 1114, 2024.

[7] N. Han, J. He, L. Shi, M. Zhang, J. Zheng and Y. Fan, "Identification of biomarkers in nonalcoholic fatty liver disease: Amachine learning method and experimental study", Frontiers in Genetics, vol. 13, pp. 1020899, 2022.

[8] Chaudhari AU, Jaini P. Zone Routing Protocol Affected By Sinkhole Attack. International Journal of Advanced Research in Computer Science and Software Engineering, ISSN. 2014 Mar;2277.

[9] T. M. Ghazal, A. U. Rehman, M. Saleem, M. Ahmad, S. Ahmad and F. Mehmood, "Intelligent model to predict early liver disease using machine learning technique", 2022 International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1-5, 2022.

[10] C. Anuradha, D. Swapna, B. Thati, V. N. Sree and S. P. Praveen, "Diagnosing for Liver Disease Prediction in Patients Using Combined Machine Learning Models," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 889-896

[11] Arpit Chaudhari, "Chronic Kidney Diseases Prediction Using Machine Learning", International Journal of Innovative Research in Computer and Communication Engineering, 2022, Volume 10, Issue 1, Pages 1111-1121

[12] Munipraveena Rela, Nagaraja Rao Suryakari and P Ramana Reddy, "Liver Tumor Segmentation and Classification: A Systematic Review?", 2020 IEEE-HYDCON, 2020.

[13] Golmei Shaheamlung, Harshpreet Kaur and Mandeep Kaur, "A Survey on machine learning techniques for the diagnosis of liver disease", 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020.

[14] Khair Ahammed, Md. Shahriare Satu, Md. Imran Khan and Md Whaiduzzaman, "Predicting Infectious State of Hepatities C Virus Affected Patients Applying Machine Learning Methods", 2020 IEEE Region 10 Symposium (TENSYMP), 2020.

[15] Yoshihiro Mitani, Robert B. Fisher, Yusuke Fujita, Yoshihiko Hamamoto and Isao Sakaida, "Effect of an Augmentation on CNNs in Classifying a Cirrhosis Liver on B-Mode Ultrasound Images", 2020 IEEE 2nd Global Conference on Life Sciences and T echnologies (LifeTech), 2020.

[16] Aman Singh, Pinku Nath, Vivek Singhal, Divya Anand, Kavita Sahil Verma and Tzung-Pei Hong, "A New Clinical Spectrum for the Assessment of Nonalcoholic Fatty Liver Disease Using Intelligent Methods", IEEE Access, vol. 8, 2020.

[17] Xiaonan Fang, Naiwen Liu, Yanling Du, Feng Yuan and Yongqin Li, "A Ten-Long Non-Coding RNA Model Improves Prognosis Prediction of Hepatocellular Carcinoma Patients", 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018.

[18] S. Mohamed, R. Ezzat, S. Ghorab, R. Bhatnagar and M. Y. Shams, "Liver Disease Identification Based on Machine Learning Algorithms," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 1062-1067

[19] Arpit Chaudhari, Prachi Jaini, "Stealthier attack on zone routing protocol in wireless sensor network", 2014 Fourth International Conference on Communication Systems and Network Technologies, pp-734-738

[20] C. A. Reddy, L. S. Kiran and V. M. Arul Xavier, "Comparative Analysis of Liver Disease Detection using Diverse Machine Learning Techniques," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1452-1457

[21] Jay Khapre Prof. Arpit Chaudhari, "Share Wheels", International Journal for Research in Applied Science & Engineering Technology (IJRASET),2024, Volume 12, Issue IV, Pages 2602-2607

[22] S. T, P. V and C. C, "Prediction of Chronic Liver Cirrhosis Using Ensemble Classification Approach," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5

[23] Prof. A. U. Chaudhari, Alkesh S. Lajurkar,"Implementing A Passive Aggressive Classifier To Detect False Information", International Journal of Advanced Research in Computer and Communication Engineering, 2023, Volume 12, Issue 4, Pages 767-774