Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

HEPATITISNET: ANALYSIS AND PREDICTION OF HEPATITIS USING MACHINE LEARNING

B. Prathyusha[1], S. Navya[2], Y. Navya[2], N. Sri Harika[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Hepatitis is one of the dangerous diseases that result from viral infections. This virus attacks the liver leading to its inflammation. Inflammation may lead to the death of the liver cells and affect the functionality of the liver. Five main types of hepatitis have been identified, namely hepatitis A, B, C, D, and E viruses. The most common types of these are hepatitis A virus, hepatitis B virus (HBV), and hepatitis C virus (HCV). Among these, HBV and HCV will cause chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma. It is estimated that 257 and 71 million people around the world are currently infected with HBV and HCV, respectively. The prevalence of HBV depends on the geographic area, and its overall prevalence was estimated at 3.6% in the world. The HCV global prevalence in adults is 2.5%. Furthermore, the incidence of HCV was estimated between 0.5% and 2.8% in various studies. According to previous studies, African and Asian countries have the highest prevalence of HBV and HCV. In Iran, the prevalence of HBV and HCV was about 2.2 and 0.5% in the general population, respectively. Prediction of chronic diseases plays an important role in health informatics. Hepatitis is one of the chronic diseases that can lead to liver cirrhosis and hepatocellular carcinoma, which cause deaths around the world. Therefore, early diagnosis is needed to control, treat, and reduce the effects of this disease.

Keywords: Hepatitis, Hepatitisnet, machine learning.

1. INTRODUCTION

Now a days the technology has been changed to machine-based learning. Machine learning technology is one among the techniques of Artificial Intelligence (AI). This technique mainly helps to assign the patterns and spontaneous exchange between the aspects or attributes which we produced. The drawback originates during these techniques are the record set having a greater number of samples with respect to same aspect values and distributions. Because of this same aspect values, it causes many unwanted problems, disorganized and ineffective record set and leads to noisiness. Using the data analysis and prediction methodology will reduce the same aspect results and maintains the data efficiency. So that we can increase the accuracy level. Our ultimate aim is to anticipate the life time of the Hepatitis patient based on their medical data. The algorithm which we are using here is Support Vector Machine and it will predict the same aspect result and noisiness to maintain the accuracy in Hepatitis level of the patient. After eliminating or avoiding the same aspect records using SVM method, we are analyzing and measuring the result alone for finding the accuracy. The solutions to maintain the accuracy is mentioned in the proposed system elaborately.

As a subset of artificial intelligence, machine learning is an umbrella term for a variety of important computational tools for early diagnosis and prognosis, which includes different classification models, such as gradient boosting, random forest, and support vector machine. Machine learning generates predictive models effectively through the detection of hidden patterns within big datasets. Given that a lot of variables could affect the clinical outcome, it is often difficult for a physician to predict a given outcome to ascertain. Machine learning algorithms could better incorporate various risk factors to identify nuanced interactions between outcomes and variables, which allows them to find new patterns between risk factors. Predicting clinical outcomes using a profiling dataset with a large

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

number of variables has drawn great interest over the past years. For instance, clinical data and microbiota based multi-omics have been used to predict outcome or severity of diseases.

2. LITERATURE SURVEY

Kashif, Ashfaq Ali, et al. [1] focuses on predicting the treatment response of a drug: "L-ornithine L-Aspartate (LOLA)" in hepatitis c patients. We have used various machine learning techniques for the prediction of treatment response, including: "K Nearest Neighbor, kStar, Naive Bayes, Random Forest, Radial Basis Function, PART, Decision Tree, OneR, Support Vector Machine and Multi-Layer Perceptron". Performance measures used to analyze the performance of used machine learning techniques include, "Accuracy, Recall, Precision, and F-Measure".

Guo, Yanhui, et al. [2] Autoregressive integrated moving average (ARIMA), support vector machine (SVM) and long short-term memory (LSTM) recurrent neural network were adopted and compared. ARIMA was implemented by python with the help of stats models. SVM was accomplished by matlab with libSVM library. LSTM was designed by ourselves with Keras, a deep learning library. To tackle the problem of overfitting caused by limited training samples, we adopted dropout and regularization strategies in our LSTM model.

Yağanoğlu, Mete, et al. [3] data discovery has been made by applying data science processes, and the HCV has been estimated with machine learning methods. By analyzing and visualizing the values in the data set, features that may be important for HCV was determined, and HCV estimation was made using various machine learning methods, pre-processing and feature extraction.

V. K. Yarasuri, G. K. Indukuri [4] comparative study between various machine learning tools and neural networks were carried out. The performance metric is based on the accuracy rate and the mean square error. The Machine Learning (ML) algorithms such as Support Vector Machines (SVM), K Nearest Neighbor (KNN) and Artificial Neural Network (ANN) were considered as the classification and prediction tools for diagnosing Hepatitis disease.

T. A, C. K, et al. [5] proposed work has compared the classification algorithms such as Random Forest (RF) Classifier, AdaBoost (AB) Classifier, Support Vector Machine (SVM) and XGBoost (XGB) algorithm for cirrhosis dataset. The performance measure, confusion matrix and accuracy score are used to compare these algorithms, and analyse the algorithm that gives better results.

Feng, Gong, et al. [6] proteomics data for analysis in this study were obtained from the Clinical Proteomics Tumor Analysis Consortium (CPTAC) database. We analyzed different proteins based on cases with or without recurrence of HCC. Survival analysis, Cox regression analysis, and area under the ROC curves (AUROC > 0.7) were used to screen for more significant differential proteins. Predictive models for HCC recurrence were developed using four machine learning algorithms.

Vijayalakshmi, C., et al. [7] Stochastic Gradient algorithm to find the Co-connection between boundaries of the date set, kernel approximation to finalise the resulting accuracy of the acute or choric prediction of patients and SVM method we use to clustering the kernel approximation calculation and connection analysis.

Kamboj, Sakshi, et al. [8] identified promising repurposed drugs viz. naftifine, butalbital (NS3), vinorelbine, epicriptine (NS3/4A), pipecuronium, trimethaphan (NS5A), olodaterol and vemurafenib (NS5B) etc. targeting HCV NS proteins. These potential repurposed drugs may prove useful in antiviral drug development against HCV.

McCandlish, John Austin, et al. [9] constructed 3 ML-based metamodels using random forest, support vector regression, and artificial neural networks and a linear regression-based metamodel from a

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

previously validated microsimulation model of the natural history hepatitis C virus (HCV) consisting of 40 input parameters. Outcomes of interest included societal costs and quality-adjusted life-years (QALYs), the incremental cost-effectiveness (ICER) of HCV treatment versus no treatment, cost-effectiveness analysis curve (CEAC), and expected value of perfect information (EVPI). We evaluated metamodel performance using root mean squared error (RMSE) and Pearson's R2 on the normalized data.

P. Idrovo-Berrezueta, D. Dutan-Sanchez, et al. [10] proposed the use a method based on CRISP-DM, where as a first procedure they apply a preparation to the data, then we prepared the dataset by cleaning the null variables, transforming the dataset by applying Hot Encoding, analysis the data with PCA (Principal Component Analysis) and using the 85% of variance, and using oversampling for the class that we have chosen. Once the dataset has been preprocessed, we proceed to use the techniques of machine learning to help evaluate if a donor's blood is qualified or not for its use. They have applied a variety of machine learning techniques such as: RandomForest, KNN (K-Nearest-Neighbor), SVM (Support Vector Machine), and a neural network ANN (Artificial Neural Network).

Mamdouh Farghaly, Heba, et al. [11] proposed framework achieved higher accuracies after SFS selection than without feature selection. Moreover, the RF classifier achieved 94.06% accuracy with a minimum learning elapsed time of 0.54 s. Finally, after adjusting the hyperparameter values of the RF classifier, the classification accuracy is improved to 94.88% using only four features.

Kaunang, Fergie Joanda, et al. [12] used different machine learning algorithms, namely K-Nearest Neighbour, Support Vector Machine, Random Forest, Neural Network, Naïve Bayes, and Logistic Regression. The performance of those different machine learning algorithms was evaluated using four different metrics, which are classification accuracy, precision, recall, and F-1 score. The classification accuracy results are 96.5%, 96.7%, 97.3%, 97.1%, 96%, 97.9% each for k-NN, SVM, RandomForest, Neural Network, Naïve Bayes and Logistic Regression.

Parisi, L., Ravichandran, N et al. [13] proposed hybrid classifier NCA-Relief-LSVM, using an ML-based syncretic FS, led to the highest classification performance (AUC = 0.97/F1-score = 97.51, and AUC = 0.94/F1-score = 94.57) and the lowest computational cost (1 and 2 epochs, 13 and 11.67 s respectively) amongst all algorithms tested on both benchmark datasets. Thus, this study strongly supports the use of ML-based syncretic FS for predicting survival in individuals affected by hepatitis.

3. PROPOSED SYSTEM

HepatitisNet is a system that utilizes machine learning techniques for the analysis and prediction of hepatitis. Hepatitis is an inflammation of the liver caused by viral infections, autoimmune diseases, alcohol abuse, or exposure to certain toxins. Early detection and accurate prediction of hepatitis can significantly improve patient outcomes and guide appropriate treatment strategies. HepatitisNet leverages machine learning algorithms to analyze various data sources and identify patterns or trends that can assist in diagnosing hepatitis and predicting its progression. These data sources can include patient demographics, medical history, laboratory test results, imaging scans, and other relevant clinical information.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

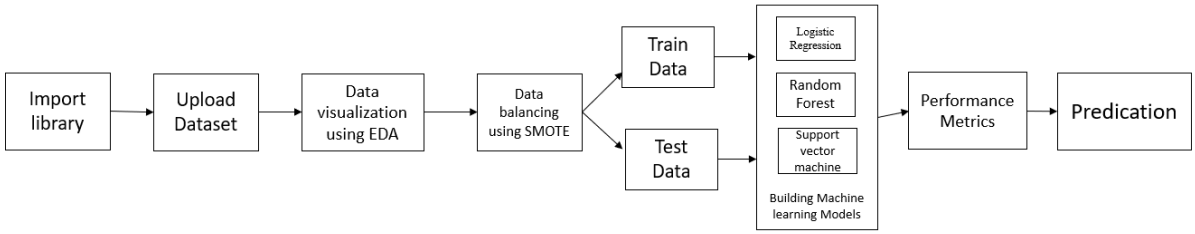


Fig. 1: Block diagram of proposed system.

The workflow of HepatitisNet typically involves the following steps:

- Data collection: It gathers a comprehensive dataset containing relevant information about patients diagnosed with hepatitis. This dataset is from research studies, or other sources.
- Data preprocessing: Before applying machine learning algorithms, the collected data needs to be pre-processed. This step involves cleaning the data, handling missing values, normalizing, or scaling numerical features, and encoding categorical variables.
- Feature selection/extraction: It identifies the most relevant features or attributes that contribute significantly to hepatitis analysis and prediction. This step helps reduce dimensionality and improve the efficiency of the machine learning models.
- Model training: Machine learning models are trained using the pre-processed dataset. Here, logistic regression, support vector machines, and random forests are employed to learn patterns and relationships within the data.
- Model evaluation: The trained models are evaluated using appropriate performance metrics to assess their predictive capabilities.
- Prediction and analysis: Once the models are deemed reliable, they can be used to predict the diagnosis, progression, or severity of hepatitis for new or incoming patients. HepatitisNet analyzes the input data and provides predictions based on the learned patterns and insights derived from the training phase.

Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

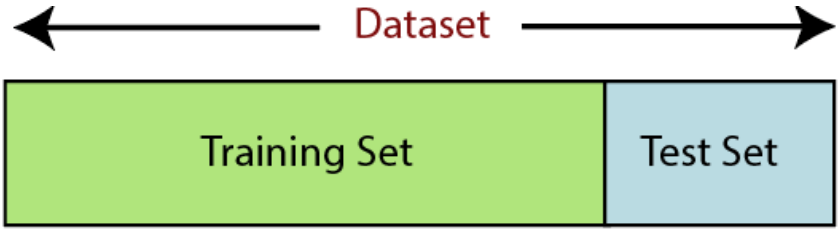
- Splitting dataset into training and test set
- Feature scaling

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:



Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.

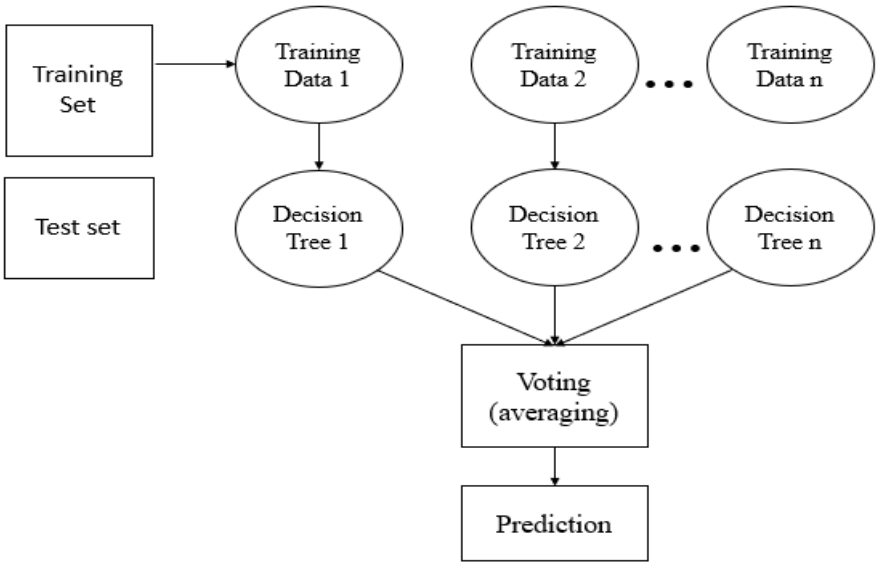


Fig. 2: Random Forest algorithm.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

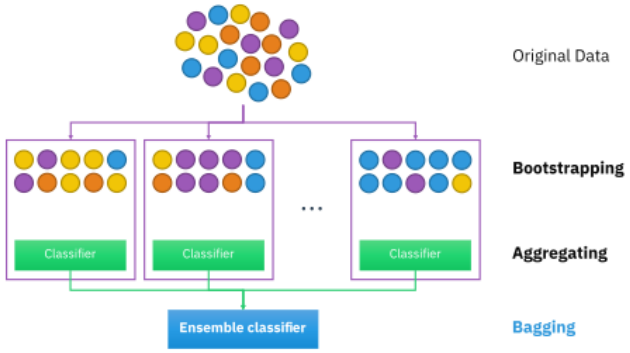


Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

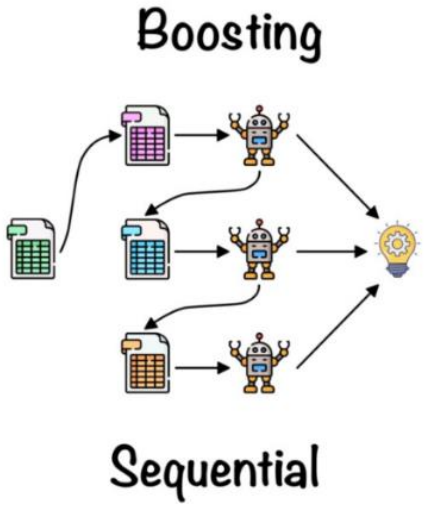


Fig. 4: Boosting RF Classifier.

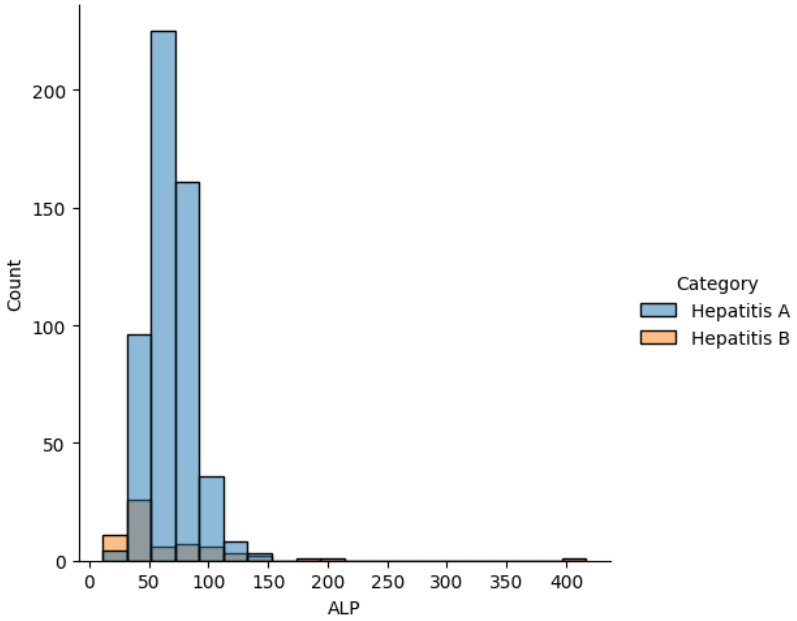
Advantages of proposed system

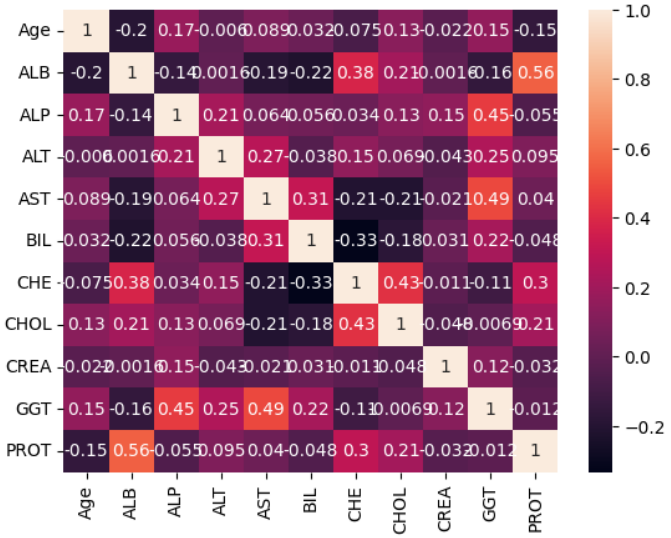
- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

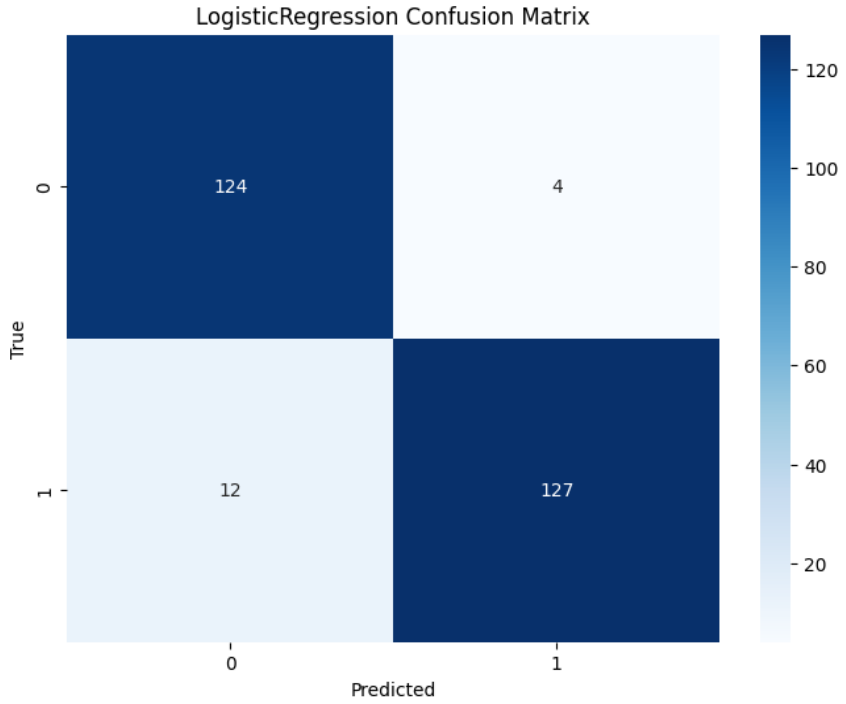
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. RESULT AND DISCUSSION

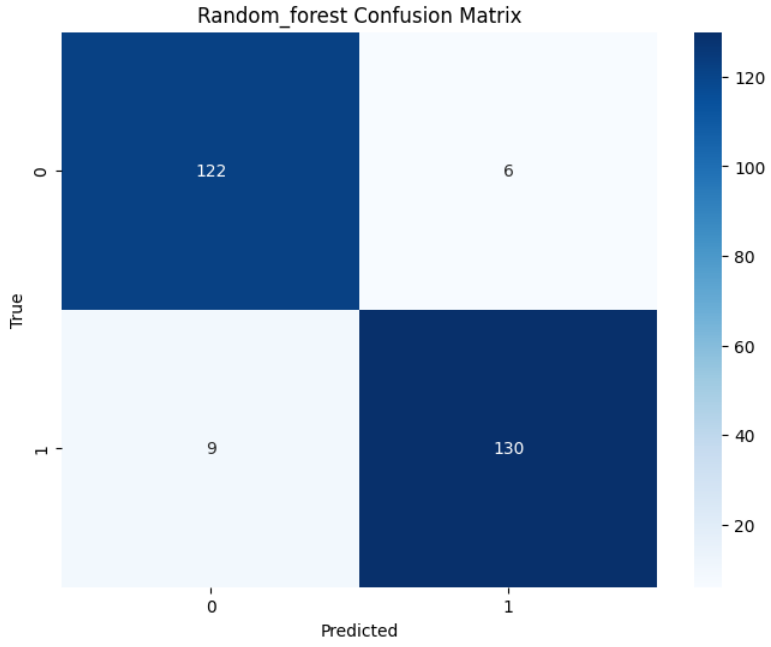




ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

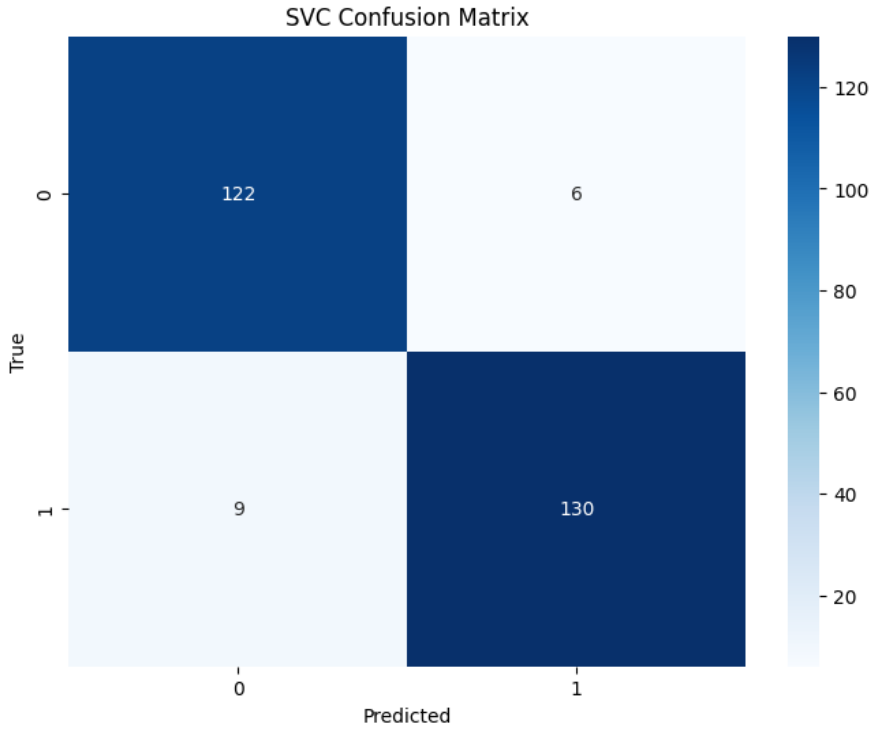


	precision	recall	f1-score	support
0	1.00	0.99	1.00	128
1	0.99	1.00	1.00	139
accuracy			1.00	267
macro avg	1.00	1.00	1.00	267
weighted avg	1.00	1.00	1.00	267



ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

	precision	recall	f1-score	support
0	0.93	0.95	0.94	128
1	0.96	0.94	0.95	139
accuracy			0.94	267
macro avg	0.94	0.94	0.94	267
weighted avg	0.94	0.94	0.94	267



	precision	recall	f1-score	support
0	0.93	0.95	0.94	128
1	0.96	0.94	0.95	139
accuracy			0.94	267
macro avg	0.94	0.94	0.94	267
weighted avg	0.94	0.94	0.94	267

5. CONCLUSION

One of the serious illnesses brought on by viral viruses was hepatitis. The liver became inflamed as a result of this virus's assault. The liver's ability to operate may been impacted by inflammation and the death of liver cells. Hepatitis a, b, c, d, and e viruses were among the five major types of hepatitis that had been discovered. Hepatitis a, b, and c viruses were the most prevalent kinds of these. (HBV). Among these, hbv and HCV would result in persistent hepatitis, liver cirrhosis, and hepatocellular carcinoma. According to estimates, there were presently 257 and 71 million hbv and hcv infections, respectively, in the world. HBV prevalence varies by location, with an approximated 3. 6% global prevalence. Adults worldwide had a 2. 5% incidence of hcv. Furthermore, different studied had

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

assessed the prevalence of hcv to range between 0. 5% and 2. 8%. The highest prevalence of HBV and HCV, according to prior studied, was found in Asian and African nations. In Iran, the prevalence of hbv and HCV in the overall population was about 2. 2 and 0. 5%, respectively. In the field of health informatics, chronic illness prediction was crucial. One of the chronic illnesses that could result in liver cirrhosis and hepatocellular carcinoma, which were fatal conditions worldwide, was hepatitis. In ordered to managed, treat, and lessen the effects of this illness, early diagnosis was required.

Future Scope

The future course of this work can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature-selection methods can be developed to get a broader perception of the significant features to increase the performance of hepatitis prediction.

REFERENCES

[1] Kashif, Ashfaq Ali, et al. "Treatment response prediction in hepatitis C patients using machine learning techniques." International Journal of Technology, Innovation and Management (IJTIM) 1.2 (2021): 79-89.

[2] Guo, Yanhui, et al. "Prediction of hepatitis E using machine learning models." Plos one 15.9 (2020): e0237750.

[3] Yağanoğlu, Mete. "Hepatitis C virus data analysis and prediction using machine learning." Data & Knowledge Engineering 142 (2022): 102087.

[4] V. K. Yarasuri, G. K. Indukuri and A. K. Nair, "Prediction of Hepatitis Disease Using Machine Learning Technique," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 265-269, doi: 10.1109/I-SMAC47947.2019.9032585.

[5] T. A, C. K, S. J and N. R, "Predictive Analysis for Hepatitis and Cirrhosis Liver Disease using Machine Learning Algorithms," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 873-877, doi: 10.1109/ICESC54411.2022.9885411.

[6] Feng, Gong, et al. "Machine learning algorithms based on proteomic data mining accurately predicting the recurrence of hepatitis B- related hepatocellular carcinoma." Journal of Gastroenterology and Hepatology 37.11 (2022): 2145-2153.

[7] Vijayalakshmi, C., and S. Pakkir Mohideen. "Predicting Hepatitis B to be acute or chronic in an infected person using machine learning algorithm." Advances in Engineering Software 172 (2022): 103179.

[8] Kamboj, Sakshi, et al. "Targeting non-structural proteins of Hepatitis C virus for predicting repurposed drugs using QSAR and machine learning approaches." Computational and Structural Biotechnology Journal 20 (2022): 3422-3438.

[9] McCandlish, John Austin, Turgay Ayer, and Jagpreet Chhatwal. "Cost-Effectiveness and Value-of-Information Analysis Using Machine Learning–Based Metamodeling: A Case of Hepatitis C Treatment." Medical Decision Making 43.1 (2023): 68-77.

[10] P. Idrovo-Berrezueta, D. Dutan-Sanchez, R. Hurtado-Ortiz and V. Robles-Bykbaev, "Data Analysis Architecture using Techniques of Machine Learning for the Prediction of the Quality of Blood Fonations against the Hepatitis C Virus," 2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2022, pp. 1-7, doi: 10.1109/ROPEC55836.2022.10018741.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

[11] Mamdouh Farghaly, Heba, Mahmoud Y. Shams, and Tarek Abd El-Hafeez. "Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt." Knowledge and Information Systems (2023): 1-23.

[12] Kaunang, Fergie Joanda. "A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms." 8ISC Proceedings: Technology (2022): 33-42.

[13] Parisi, L., Ravichandran, N. Syncretic Feature Selection for Machine Learning-Aided Prognostics of Hepatitis. Neural Process Lett 54, 1009–1033 (2022). https://doi.org/10.1007/s11063-021-10668-7