Machines and Algorithms
*http://www.knovell.org/mna*

*Research Article*

# Ensemble learning model for Classification of Hepatitis C Disease

**Sara Ashraf [1], Fatima Bukhari [1,*], Naeem Aslam[1] and Humera Batool Gill [3]**

[1]Department of Computer Science, NFC-IET, Multan, 60000, Pakistan

[2] Institute of Computer Science & IT, The Women University, Multan, 60000, Pakistan

[*]Corresponding Author: Fatima Bukhari. Email: fbukhari43@gmail.com

Received: 07 June 2023; Revised: 27 June 2023; Accepted: 28 July 2023; Published: 31 October 2023

AID: 002-03-000026

**Abstract:** Supervised machine learning is gaining prominence in bioinformatics, particularly in the context of disease diagnosis. This discipline falls under the broader umbrella of artificial intelligence (AI). Hepatitis disease is a leading cause of death, with Hepatitis C being particularly concerning due to the absence of a vaccine. The transmission of Hepatitis C primarily occurs through blood transfusions, contaminated needles, and unsterilized medical instruments. Accurate diagnosis and prediction of Hepatitis C virus (HCV) infection are crucial for effective treatment of affected individuals. Traditional clinical approaches may lead to misdiagnosis in hepatitis cases. Machine learning technologies are enhancing the healthcare sector by improving the accuracy of disease diagnosis and prognosis. This research introduces a hybrid ensemble model aimed at predicting and classifying data related to HCV patients. The dataset utilized, known as HCV+data, is sourced from the UCI machine learning repository. Four classification algorithms such as logistic regression, support vector machine, decision tree, and K-nearest neighbour were employed in the training process. A hybrid ensemble model is created using the majority voting method to integrate various weak or base classification learners. Results demonstrate that the ensemble learning model achieves superior accuracy compared to single-learner machine learning algorithms, with a classification accuracy of 94.07% for hepatitis patients. This model is expected to assist healthcare professionals in accurately diagnosing complex and progressive diseases.

**Keywords:** Machine learning; Artificial Intelligence, Hepatitis; Ensemble model; Support vector machine Logistic regression; Decision Tree; K-nearest neighbour;

## 1. Introduction

Hepatitis is a medical condition characterized by inflammation of the liver. The liver plays a crucial role in cleansing the blood, digesting food, and protecting the body from infections. If the liver becomes damaged, it may not function properly. Hepatitis can be triggered by various factors, including excessive alcohol consumption, exposure to environmental chemicals or medications, and certain medical conditions [1].

Hepatitis A, B, and C are the three most common types of viral hepatitis. Hepatitis A has the highest incidence, followed by hepatitis B, and then hepatitis C. Hepatitis C is the most dangerous type of infection. Although there is a vaccine for hepatitis B, it remains unaffordable for many people with low incomes. Different antiviral medicines [2] are used to treat hepatitis C but currently, there is no vaccine for hepatitis

C [3] anywhere in the world. This makes it crucial to take proactive measures and disseminate information about the disease [1].

Patients with chronic hepatitis, which occurs when the virus remains in the body for more than six months, maybe prescribed antiviral medicine by their physicians. This is because chronic hepatitis can cause serious health complications. In the United States, the conditions that affect most people are stroke, diabetes, heart disease, cancer, hepatitis C, and osteoarthritis [4].

In the field of health informatics, the accurate prediction of chronic disease progression is absolutely essential. Chronic diseases can have long-lasting effects, even after treatment, making early and precise diagnosis crucial. This early detection not only leads to advances in disease prevention but also significantly improves the overall efficacy of therapy [4]. Machine learning, a specialized area of artificial intelligence, uses statistical models to make predictions and allows software to "learn" without explicit programming. To create accurate estimates of future output values, machine learning algorithms rely on historical data as input [5]. Feature selection is also critical for producing a more concise and critical description of data by eliminating redundant and unnecessary features.

### *1.1. Research Objectives*

The death rate among Hepatitis patients has been increasing globally. There is a lack of resources and healthcare services for HCV patients, leading to Hepatitis becoming a chronic condition. Early detection and prediction of the disease can save many lives. Previous studies [6] have used single classifiers to classify diseases, but there is a risk of misclassification. To improve prediction accuracy, it is essential to build a model that can forecast diseases and enhance the efficiency of medical treatment. The goal of this study is to analyze different machine learning algorithms and develop a hybrid machine learning model to achieve higher accuracy compared to a single learning prediction model.

RO1: Classify the patients with Hepatitis C Disease by using supervised machine learning classifiers.

RO2: To Perform a comparison analysis between the supervised ML classification algorithms individually and ensemble ML model’s performance in terms of prediction accuracy.

RO3: Develop an ensemble Model to classify blood donors and hepatitis C virus infected patients in terms of classification accuracy.

### *1.2. Research Questions*

RQ1: How can classification algorithms predict Hepatitis Disease?

RQ2: How does the performance of ensemble models stack up against traditional machine learning models (base Classification Algorithm) when it comes to predicting hepatitis

RQ3: How competently our purposed hybrid ensemble model will be fit in terms of accuracy?

This study's main contribution is to provide comprehensive information about the Hepatitis C disease, data preparation, and the prediction and classification of HCV patients using various machine learning algorithms. A hybrid ensemble model was developed to significantly improve classification accuracy. Data preprocessing, data cleansing, and univariate feature selection strategies were applied to achieve superior results. Four machine learning classifiers - Decision Tree, Support Vector Machine, Logistic Regression, and KNN - were rigorously tested on a publicly available dataset from the UCI machine learning repository. The performance of these algorithms was evaluated using a confusion matrix. Additionally, a hybrid ensemble model was developed by combining all the classifiers mentioned above, and its classification accuracy was extensively evaluated. The study presented a comprehensive performance comparison analysis of single algorithms and the hybrid ensemble model using a confusion matrix. The implementation of these classifiers involved using the Python language with sci-kit learn, Seaborn library, and Pandas profiling for finding the correlation between variables. The remaining sections of this paper are laid out as follows:In the related work section, we will review the relevant literature on the use of machine learning for Hepatitis disease. The methods and materials section will explain the methodology behind the

experiments, including all relevant data. In the results and discussion section, we will discuss the experimental results. Finally, in the last section, we will present the conclusion and discuss future directions."

## 2. Related work

It has been accounted for that hepatitis brings about millions of deaths every year. The Prognosis of hepatitis by traditional techniques is hectic work and needs, costly clinical examination [7].

Hepatitis, also known as hepatitis A and hepatitis B, is soreness of the liver typically brought on by a virus. Clinicians can determine whether or not a patient has hepatitis by examining a collection of datasets and using supervised data mining techniques [8].

In the past different machine learning models have been developed by different researchers.

In [9], the writer used a machine learning model, a support vector machine for hepatitis patient's diagnosis. The dataset was taken from, the UCI Machine learning repository having 155 patients record. The wrapper method is a feature selection technique that was used in this study for achieving higher accuracy. Performance of SVM Classifier, without feature selection and with feature selection was evaluated. SVM extract 74.5% performance accuracy after feature selection. While in the other research, data mining & machine learning techniques have been utilized for the prediction of hepatitis disease [10]. Different Machine learning models like KNN, NB, SVM, RF and Multi- Layer Perceptron had trained. The same dataset [9] of hepatitis patients has been used in this article. For finding out highly correlated features their proposed model used a feature selection procedure named Info-gain related to ranker search Method. Comparison of models had been evaluated on the ratio of following parameters F1-score, Recall, Precision and ROC graph. The accuracy achieved by SVM was 91.14% while RF was 92.41%.

In [11] numerous hepatitis disease diagnosis techniques have been discovered using data mining techniques. These methods were primarily created utilizing single- learning techniques. Additionally, these techniques prevent the facts from being learned collectively. Combining the outcomes of many predictors in classification problems can increase accuracy. This work aims to use the benefits of ensemble learning to come up with an accurate way to diagnose hepatitis. Researchers used groups of Neuro- Fuzzy Inference System, Self- Organizing Maps, and Non-linear Iterative Partial Least Squares to put all the data together and predict hepatitis disease. Data collection, used in this experiment was taken from UCI repository. They also use decision trees to determine which parts of the experimental dataset are the most important. They apply our methodology to a collection of data gathered from the actual world and then evaluate the outcomes in light of the most recent information from other research. After looking at the dataset, they found that their method works much better than the KNN, the ANFIS and the SVM. This method scored 93.06 percent accuracy.

In [12] researchers had been evaluated kinds of ML models. NB, KNN and SVM were used. These classifiers were utilized for the purpose of classification and prediction of data segmentation tools for the purpose of hepatitis illness detection and diagnosis. The dataset for this analysis was available at UCI ML repository. Implementation of these classifiers was done by using MATLAB software. For choosing the most accurate classifier matrices such as accuracy and mean square error were considered. The Naïve Bayes Algorithm was predicting better accuracy of 87% and low mean square error. While in another study [13] three-step approach was used to achieve the results. In the first part of the study, the 13 and 19 accessible dimensions in the datasets for heart disease and hepatitis illness are reduced using the C4.5 decision tree approach, which is part of the CBA software. In the second step, fuzzy pre-processing is used to add weights for heart disease and hepatitis datasets. The first step is to normalize the datasets within the range [0, 1]. These two phases happen at the same time, in the third step of the classification process. They looked at the classification accuracy, sensitivity, and specificity scores, as well as the confusion matrix, of the suggested method to see how well it worked. When the training and testing stages are split, the system can correctly classify 92.59 percent of heart disease datasets and 81.82 percent of hepatitis datasets.

In [14] researcher utilized data of Egyptian patients on liver fibrosis. This dataset has 1385 patients' records. In this dataset was accessed from the UCI repository. In this, different classifiers like KNN, support vector machine, RF, Naïve Bayes, Logistic Regression, Decision Tree and Gradient Boosting were used to sort the data. Additionally, Data balancing, SMOTE, and different feature selection methods were applied. Feature selection carries out in WEKA software. For attribute evaluation, different filter- based methods like Chi-Square, Info Gain, Gain Ratio, and Relief F were used. KNN consider the best model because it accomplished the higher ratio in the different matrices AUROC, Accuracy etc. The best accuracy achieved by the author's model is 94.40%. In this study [15], the author identifies the type and phase of hepatitis disease. They trained SVM and ANN to diagnosis. Various models like SVM, LVQ, GRNN and RBF were utilized in this examination. Dataset was collected from the 2 major hospitals in Mashhad, Iran, having 250 suspected persons. In this researchers differentiate what type of hepatitis has or the person was affected by hepatitis disease or not. The performance of each classifier was compared and analyzed that RBF performs more accurately as compared to other networks. In general accuracy of the diagnosis was approximately 96.4%.

W. Ahmad et al. A hybrid intelligent methodology was produced by combining (ANFIS) and an informative strategy. This methodology was proposed to diagnose a hepatitis disorder that can result in death. The dataset on hepatitis obtained from the ML repository UCI, Data was pre- processed to make it useable before mining. Following the completion of the pre-processing stage, the information gain methodology was applied to considerably cut down on the number of characteristics required to be computed After that, the selected elements were categorized using the ANFIS categorization system. Statistical methodologies were implemented so that the effectiveness of the proposed process could be evaluated. The proposed method achieved the most excellent overall scores in classification accuracy, specificity, and sensitivity analysis, with respective percentages of 95.24 percent, 91.7 percent, and 96.17 percent [7].

In [16] hybrid machine learning approach is used. Clustering and classification are two methods that are often used to start the first stage. These are examples of how the data can be pre-processed. Usually, the second step is based on the results of the first step. To do this, decision trees were used as a classification method, logistic regression as a clustering method, and neural networks as a clustering method. The results of an experiment performed on a dataset taken from the real-world show that the hybrid classification- on-classification technique performs at the highest level. The two-stage decision tree combination produced the best accuracy rate (99.73%) and the fewest Type I and Type II mistakes (0.22 percent and 0.43 percent). This study contributes by proving that hybrid machine learning algorithms outperform standalone ones.

In [17] Researchers used DT, LR, NB and SVM to compare and describe results. Based on the outcomes produced using the SciPy package and the Python language. Logical regression was found to be the method that was accurate 87.17% of the time, according to the study. On the other hand, the algorithm known as the Decision Tree has been shown to have an accuracy of 82.05 percent, making it the method with the second-highest level of precision. The Linear Support Vector Machine, which has the best accuracy at 76.92 percent, is the next algorithm after the Decision Tree Algorithm followed by the NB extracts with 76.92 percent accuracy. In another article [18], investigation of the liver fat of the 36,703 people who took part in the UK Biobank, the first step had been to create an ML technique that might permit precise quantification using abdominal MRI raw data. A Selection of 4,511 subjects whose liver fat had already been determined by Perspectum diagnostics was used to process. The datasets that were employed in the last testing had correlation values of 0.97 and 0.99, and the errors for the two stages were 0.50 and 0.41 percent. It was shown that a method that directly used imaging data rather than merely using clinical data to estimate the amount of fat in the liver was an approach that was much more accurate.

In this [19], included a total of 2235 CHB (chronic hepatitis B) patients. The endpoint was a lack of HBsAg detectability using ECL kits (also known as HBsAg seroclearance).106 CHB patients lost HBsAg antibodies. Dataset had been segmented, 1564 rows for training data and 671 rows as a testing data, with the training data accounting for 70 percent of the total. They developed a model based on 4 ML techniques, RF, LR, DT and XGBoost. AUCs for RF, LR, DT, and XGBoost near to 0.619, 0.829, 0.891 and confidence

interval: 0.677 to 0.683. XGBoost had the AUC overall. The final times for XGBoost, RF, DCT, and LR were all 0.97. For the comparison of previous ML literature ref to Table 1.

**Table 1:** Comparison ML related work

| Ref | Year | Classifiers | Datasets | Findings | Limitations | Results |
|---|---|---|---|---|---|---|
| **[20]** | 2023 | Naive Bayes | 142 records from UCI dataset | This article identifies age and medical history as important determinants impacting treatment decisions and patient outcomes, but also notes limitations in attribute independence for greater healthcare accuracy. | This study used a small dataset (142) instances with only one classifier, no algorithms comparison. No preprocessing details which resulted in lower predicted accuracy. | 86.04% |
| **[21]** | 2023 | Logistic Regression, KNN RF SVM NB | 155 instances UCI | This revealed the significance of missing value datasets and feature selection techniques in boosting classification model accuracy and reliability, potentially leading to enhanced decision-making in a range of domains. | A feature selection approach was not used in this investigation. Moreover, this article employed a smaller dataset. Limited hyper parameter optimization, potentially affecting model performance. | 93.18 % highest accuracy |
| **[22]** | 2022 | Decision tree and KNN | 44 instances | In this, the author uses the KNN algorithm for hepatitis C prediction which had been Optimizing enhance efficiency | This research had resource restrictions. Smaller number of samples had been used. | 0.42% |
| **[23]** | 2021 | Support vector machine, NB | Covid-19 dataset | This article compares the accuracy of two machine learning approaches, SVM and Naïve Bayes, before and after selecting the features. The chi-square feature selection approach was implemented. | The accuracy of the SVM classifier had been reduced followed by these feature selection methods. | SVM-83.86% NB-87.09% |
| **[24]** | 2021 | NB, LR, DT, RF, KNN, SVM | Dataset from Kaggle | This research demonstrates the effectiveness of | Extremely unbalanced datasets were utilized. The primary drawback | Highest accuracy |

| | | | |
|---|---|---|---|
| | various machine learning algorithms in accurately predicting stroke based on distinct physiological factors. | of this research was that they utilized a textual dataset compared to a real-time brain imaging dataset. | NB-82% |

## 3. Methodology

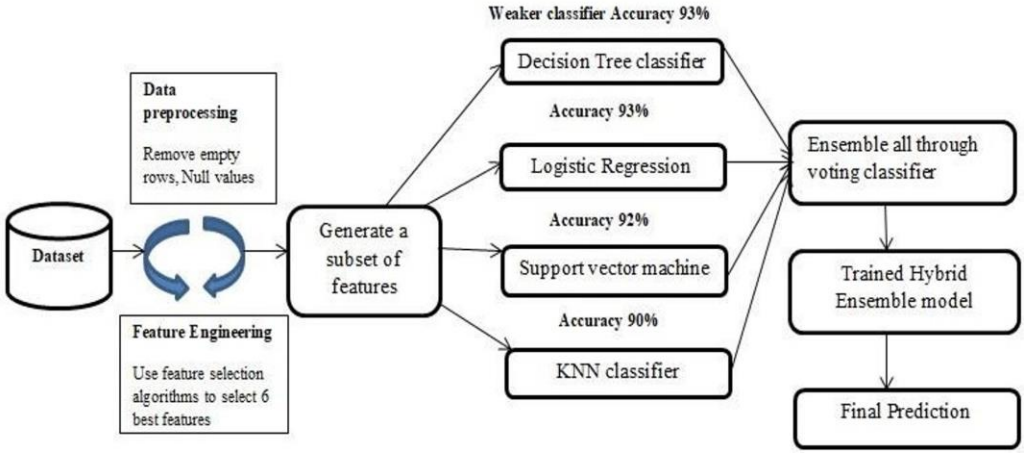


**Figure 1:** Methodology of the proposed model

This paper aims to develop a model for categorizing people suffering from hepatitis disease. To accurately forecast hepatitis disease, we utilized machine learning (ML) algorithms such as Support Vector Machine (SVM), Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN) classifiers. Before the classification task, we performed data preprocessing and data cleansing. Python language was used for implementing these classifiers. For feature selection, we employed the Univariate selection method and identified the top nine features from the dataset. We also created a correlation matrix heat map to analyze the Univariate feature extraction. Additionally, data visualization and statistical analysis were carried out on the dataset. Proper preprocessing of the hepatitis disease patient dataset was performed to ensure there were no missing values and noisy data. Subsequently, several machine learning models, as well as a hybrid ensemble, were trained including SVM, Decision Tree, Logistic Regression, and KNN classifiers, among others, for making predictions. The proposed model overview is illustrated in Figure 1.

The various parts of our proposed system include:

- Data Collection & Description
- Preprocessing
- Feature Engineering
- Algorithm Discussion
- The proposed ensemble Model Learning

### *3.1. Data collection & Description*

We gathered this dataset from the UCI Dataset Repository; named HCV+data, this hepatitis disease dataset contains 13 hepatitis features (X (Patient ID/No), Category (binary 0,1), Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT). This dataset [25] contains 615 samples of hepatitis disease patients.

**Table 2:** Description of dataset attributes

| Features | Data Type | Description |
|---|---|---|
| Category | Binary (0,1) | Label |
| Age | Numerical | Attribute |
| Sex | Binary (0,1) | Attribute |
| Choline esterase (CHE) | Numerical | Attribute |
| Alkaline Phosphatase (ALP) | Numerical | Attribute |
| Alanine Transaminase (ALT) | Numerical | Attribute |
| Aspartate Aminotransferase (AST) | Numerical | Attribute |
| Bilirubin (BIL) | Numerical | Attribute |
| Albumin Blood (ALB) | Numerical | Attribute |
| Cholesterol (CHOL) | Numerical | Attribute |
| Creatine (CREA) | Numerical | Attribute |
| Gamma-glutamyl Transferase (GGT) | Numerical | Attribute |

In Table 2, comprises various columns, including age and gamma-glutamyl transferase (GGT) as the independent variables, and the category column as the dependent variable. The data will undergo thorough analysis using machine learning techniques.

### *3.2. Preprocessing*

In the preprocessing stage, which is an essential part of the procedure, we used an algorithm to exclude normal patients from the health examination data and include only patients with Hepatitis. We removed variables that have missing values and are planning to use an algorithm to fill in any missing values we find. We employed a variety of machine learning techniques and compared the accuracy rates of each of these algorithms.

### *3.3. Data splitting*

We utilized the Scikit-learn library to divide the data, employing the train_test_split function. The data has been confidently split into 80:20 ratios.

### *3.4. Feature Selection/Engineering*

When developing a reliable model, it is important to utilize various machine learning techniques to reduce the number of input variables. This is necessary due to the extensive nature of the hepatitis disease dataset. Doing so helps to decrease the computational requirements of the model and improve its functionality. We have illustrated the advancements made through the use of a correlation matrix and a Heat Map, as well as through the application of Univariate feature selection methods.

#### *3.4.1. Univariate Selection*

This test can be done mathematically or statistically to determine well-known characteristics that have the potential to have the most effective relationship toward performance variables. Specifically, the test is used to analyze which variables have the most effective relationship. The Select Best class found in the Scikit package has been used. This class selects a preset number of the most helpful qualities from a given dataset. A wide array of statistical tests, employing several methodologies, have been carried out by Select Best. After implementing the Univariate selection process, the following features have been extracted: GGT,

AST, BIL, ALT, CREA, ALP, ALB, CHE, and Age. Extracted features and their scores are presented in Figure 2 below.

| | Specs | Score |
|---|---|---|
| 10 | GGT | 11931.274568 |
| 5 | AST | 8179.105101 |
| 6 | BIL | 5707.235379 |
| 4 | ALT | 1861.780434 |
| 9 | CREA | 1700.214095 |
| 3 | ALP | 750.742309 |
| 2 | ALB | 110.200347 |
| 7 | CHE | 71.985583 |
| 0 | Age | 60.388018 |

**Figure 2:** Extracted Features

*3.4.2. Correlation Matrix with Heat map*

Correlation indicates how attributes are related to the target attribute and to each other. When the value of the target variable increases as the feature values increase, it's called a positive correlation. Conversely, a negative correlation exists when the target variable value decreases as the feature values decrease. A heat map simplifies the identification of the dataset's characteristics that are closely related to the target characteristic. We used the Seaborn Library to plot the associated characteristics on the heat map. For analyzing correlation with heat map refer to figure 3 below.

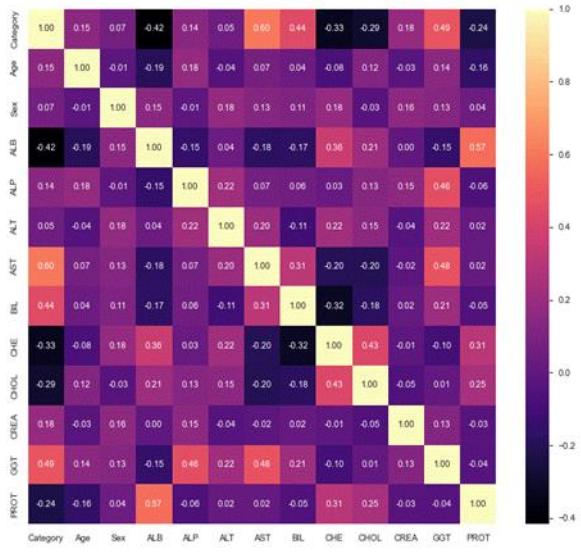


**Figure 3:** Correlation with Heat map

## 4. Algorithms Discussion

### *4.1. Support Vector Machine*

In machine learning, Support Vector Machine (SVM) is used to identify a hyperplane in a given N-dimensional dataset. SVM helps to uniquely differentiate the data points in space and can be used to separate multiple classes by using multiple hyperplanes. The decision boundaries of these hyperplanes are used to classify data points of different classes. For our experiment, we utilized the Scikit library for Support Vector Machine and achieved an accuracy score of 92.37% using the linear kernel.

### *4.2. Decision Tree*

In machine learning, a decision tree is a non-parametric algorithm used for classification and regression problems. It has a hierarchical tree structure with root nodes, internal nodes, and child nodes. The strategy used in the decision tree is divide and conquer. We utilized the Scikit library and the DecisionTreeClassifier header file for implementing the decision tree. Our model achieved an accuracy score of 93.22%.

### *4.3. Logistical Regression*

In machine learning algorithms, Logistic Regression (LR) [26] is a linear model used for allocating records to multiple or binary classes with a discrete set. Logistic regression has been employed to solve classification problems such as distinguishing between Hepatitis patients and non-patients. It is also known as an algorithm for predictive analysis. We utilized the Scikit library to perform logistic regression and, with the assistance of the Logistic Regression header file, achieved an accuracy score of 93.22%.

### *4.4. K-Nearest Neighbors*

The K-Nearest Neighbour classifier is a nonparametric instance-based algorithm [27]. This algorithm works based on supervised learning. In this algorithm, new cases are grouped based on similarity and the distance is measured using a distance matrix [28]. The algorithm identifies commonalities between previous and new datasets. K-NN algorithm stores all current data for one or multiple categories and classifies upcoming new records based on their similarity to a specific category. For K-Nearest Neighbors, we utilized the Sci kit library and Gaussian header file, achieving an accuracy of 90.68%.

## 5. What is Ensemble Learning?

Ensemble learning is a form of machine learning which brings together predictions from numerous algorithms to improve accuracy [29]. In machine learning, ensemble means combining homogeneous weak machine learning algorithms to make a strong predictor model. The implementation of ensemble models intends to lower conversion error. This approach optimizes model prediction efficiency when base models differ significantly and are independent [29]. Noise, variations, and bias are the most common causes of inconsistencies between actual and expected outcomes when using machine learning methods to estimate a target variable. Ensemble approaches use multiple algorithms for machine learning to provide accurate forecasts than a single classifier [30]. The four fundamental categories of ensemble learning algorithms (bagging, boosting, stacking, and voting) are essential for effective modeling of predictions. Here ensemble algorithms learn from a complete the training set or either part of training set [31]. Figure 4. Ref. to generalized overview of ensemble learning model.

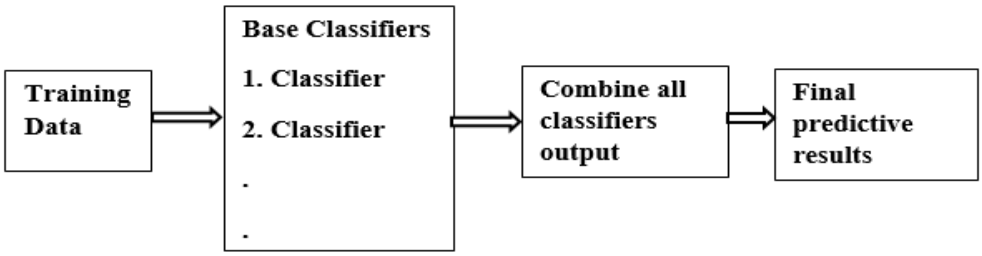


**Figure 4:** Generalized overview of ensemble learning model

- ***Bagging***

Bagging is known as Bootstrapping. Bootstrapping, similar to random replacement sampling, can help determine variance as well as bias in a dataset [32]. The term bagging is referring to the reduction in the prediction variance by making an extra dataset for learning with combined repetitions. This is done to generate multiple set of the original dataset.

- ***Boosting***

Boosting algorithms work on weighted average patterns to convert weak learner classifiers into strong ones. The term Boosting means to readjust the weights of all last classified dataset. In this process the original dataset is divided into different subsets to train the classifiers [33]

- ***Stacking***

Stacking is a learning-based technique that combines the fundamental outputs. When the ultimate selection of features is a linear framework, the stacking is commonly referred to as "model blending" or generally "blending"[33]. In stacking heterogeneous weak/base learners, trained and combined the results based on different weak models' predictions by using a voting classifier.

- ***Voting***

A voting classifier is an ensemble learning method, which aggregates the predictions from multiple independent models to generate final outcomes. In voting method major two types of voting exist hard voting & soft voting. The final prediction is generating by collective estimated probabilities from all base classifiers and select that class with the greatest average probability [34]. Voting approach in ensemble learning is one of the best approaches in previous studies [35].

### *5.1 Proposed Hybrid Ensemble Model*

In the first step, we trained several machines learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, and K-Nearest Neighbor (weak learners). We utilized different variations of each model to create a robust machine-learning algorithm. The term "hybrid ensemble" indicates that we used a combination of diverse weak machine learning algorithms. In Python, we utilized a majority voting classifier to generate variations of specific models. The proposed system improves upon existing models by combining base learners into a model that makes predictions through a majority vote using a hard majority classifier. For evaluation, we employed the majority Voting Classification process to consider the class mostly predicted by the weak learners as the final predicted class by the ensemble model. The algorithms' performance was assessed using popular evaluation metrics: accuracy, precision, recall, and F1 score.

## 6. Performance Evaluation through Confusion Matrix

Here we have used confusion matrix to present the performance of a weak machine learning algorithm.

1. (A) Accuracy

Factor, which determines the predictions accuracy, is known as algorithm accuracy.

$$\frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

In equation (1), TP, TN, FN,and FP show the number of TruePositive Negatives, True Negatives, False Negatives and False Positives.

2. (P) Precision

Factor, which used to find out the measure of classifier's [36], [37] the equation of precision is illustrated in Eq. (2)

$$\frac{TP}{TP+FP} \tag{2}$$

3. R) Recall

Factor, which used to find out the completeness and sensitivity of the algorithm, is known as recall of algorithm. The equation of Recall is illustrated in Eq. (3)

$$\frac{TP}{TP+FN} \tag{3}$$

4. F1-score

Factor which defined the precision and recall in the form of weightedaverage [36], [38]. The equation of F1-score is illustrated in Eq. (4)

$$\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} \tag{4}$$

## 7. Experimental Results

In this paper, different Machine learning algorithms such as support Vector Machine, Logistic Regression, Decision Tree, and KNN have been used.

5. Why uses these classifiers?

An ensemble model can be constructed by utilizing the power of two or more classifiers. Different Machine learning, data mining and deep learning classifiers can be used for classifying and predicting of diseases. The classifiers which are used in this study called as base/weak classifiers [35]. By combining these classifiers, we can make strong predictive model which can predict /classify data more efficiently as compared to single learning classifiers. There is a chance of miss classification if we use single learning technique, but by ensemble model there is a less chance of miss classification. Results demonstrate that SVM scored 92 %, Logistic Regression 93 %, Decision Tree 93 % and KNN scored 93% accuracy. Logistic Regression & Decision Tree achieved the same percentage of accuracy. The dataset is taken from the UCI ML repository and then preprocessing has been done on it. Divided the whole set of data: a training set and a testing set. A hybrid ensemble model for the Prediction & Classification of HCV patient's data is developed and implemented. This model extracted 94.07 % accuracy which was higher than all mentioned Machine learning algorithms. For implementing all these classifiers python languages with sci-kit learn, Seaborn library is used. Pandas profiling is used for finding a correlation between variables. The performance of all of these is evaluated through confusion matrix based on the following parameters: Accuracy, Precision, Recall and F1- Score.

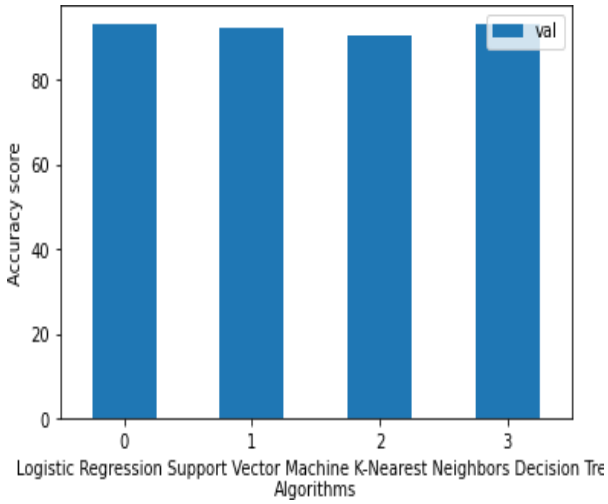


**Figure 5:** Graphical representations of all Algorithms

In fig. 5 graphical representation of machine learning algorithms are presented Support vector machine achieved 92% accuracy, Decision Tree 93%, Logistic Regression 93% and KNN 90%.

**Table 3:** Accuracy Comparison

| Algorithms | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **SVM** | 92% | 0.92% | 0.92% | 0.92% |
| **DT** | 93% | 0.93% | 0.93% | 0.93% |
| **LR** | 93% | 0.90% | 0.93% | 0.92% |
| **KNN** | 90% | 0.89% | 0.91% | 0.89% |
| **Proposed EM** | 94% | 0.92% | 0.94% | 0.93% |

Table 3 illustrate the accuracy comparison of all machine learning algorithms and the proposed hybrid ensemble model.

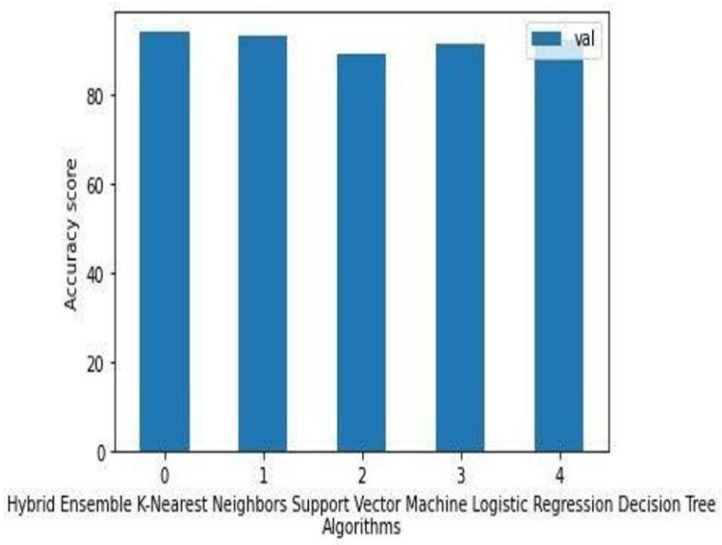


**Figure 6:** Graphical representations of Hybrid with of all algorithms

Lastly in Fig.6. classification comparison of proposed model and all individual machines learning models. Performances are presented in graphical form on the Hepatitis dataset. The prediction rate of the proposed hybrid ensemble model is 94.07% higher than all weak machine learning algorithms.

## 8. Conclusion

Healthcare sector has need to rapid improvement in disease diagnosis and prognosis. Detecting and diagnosing the deadly diseases at early stages is a big challenge now days. There is keen need to construct a model which can predict disease more accurately as compared to traditional treatment methods. The main focus of this research is to classify the hepatitis by analyzing different Machine learning classifiers. For this purpose, four machine learning algorithms are applied for the Classification of Hepatitis C patients. The dataset that has been used in this is publicly available at UCI. This research evaluates the classification algorithm's performance on hepatitis disease patients by using Python language and improves the accuracy. It discovers that the individual model is providing accuracy of up to 93%. The proposed ensemble model is including a Support vector machine, Decision Tree, Logistic regression and KNN has been developed. The proposed model extracted 94.07% accuracy. This predictive model will help Doctors and physicians in making an accurate identification of hepatitis disease patients. We conclude by our work and from the available literature, no model is completely accurate in all aspects. There are limitations to consider, such as the reliance on the quality of the dataset from UCI. This dataset may not include broad patient demographics or may contain incomplete data. Additionally, there is the potential neglect of interactions

between variables that could provide additional insights. There can be a chance of error and miss classification by outliers etc. Accuracies of algorithms may vary on different datasets due to the number of records, decision parameters and so on. Diversity of ensemble models is available that extract different levels of accuracy on the different types of data.so it is noted that no model is completely efficient and accurate in all aspect. However, the implications are significant, as this approach offers a more efficient and accurate diagnostic method for early detection of hepatitis C. This could improve patient outcomes through timely and targeted interventions and drive further innovation in the application of machine learning in medical diagnostics and healthcare.

## 9. Future Work

In future studies, there is still a need to improve the accuracy of this model. Furthermore, we will train this model in the prediction of various medical datasets. Dataset used in this study is small due to limitations on available resources. Moreover, we will also improve the model's performance by using a larger dataset. We will also implement deep learning and data mining algorithms.

## References

[1] A. Orooji and F. Kermani, “Machine Learning Based Methods for Handling Imbalanced Data inHepatitis Diagnosis,” *Front. Heal. Informatics*, vol. 10, no. 1, p. 57, 2021, doi: 10.30699/fhi.v10i1.259.

[2] NHS.(n.d.).https://www.nhs.uk/conditions/hepatitisc/treatment/#:~:text=Hepatitis%20C%20medicines,for%208%20to%2012%20weeks.

[3] Hepatitis C,Mayo-clinic .(n.d.). https://www.mayoclinic.org/diseases-conditions/hepatitis-c/diagnosis-treatment/drc-20354284.

[4] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, “A predictive method for hepatitis disease diagnosis usingensembles of neuro-fuzzy technique,” *J. Infect. Public Health*, vol. 12, no.1, pp. 13–20, 2019, doi: 10.1016/j.jiph.2018.09.009.

[5] A. H. Roslina and A. Noraziah, “Prediction of hepatitis prognosis using support vectormachines and wrapper method,”*Proc. - 2010 7th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2010*, vol.5, no. Fskd, pp. 2209–2211, 2010, doi: 10.1109/FSKD.2010.5569542

[6] H. Hartatik, M. B. Tamam, and A. Setyanto, "Prediction for diagnosing liver disease in patients using KNN and Naï ve Bayes algorithms," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020: IEEE, pp. 1-5.

[7] W. Ahmad *et al.*, “Intelligent hepatitis diagnosis using adaptive neuro-fuzzy inference system and information gain method,” *Soft Comput.*, vol. 23, no. 21, pp. 10931– 10938, 2019, doi: 10.1007/s00500- 018-3643-6.

[8] M. Rouhani and M. M.Haghighi, “The diagnosis of hepatitis diseases by support vector machines and artificial neural networks,” *2009 Int. Assoc. Comput. Sci. Inf. Technol. Spring Conf. IACSIT-SC 2009*, pp.456–458,2009, doi: 10.1109/IACSIT-SC.2009.25.

[9] F. Penin, J. Dubuisson, F. A. Rey, D. Moradpour, and J. M. Pawlotsky, “Structural Biology ofHepatitis C Virus,” *Hepatology*, vol. 39, no. 1, pp. 5–19, 2004, doi: 10.1002/hep.20032.

[10] K. Ahammed, M. S. Satu, M.I. Khan, and M. Whaiduzzaman, “Predicting Infectious State ofHepatitis C Virus Affected Patient’s Applying Machine Learning Methods,” *2020 IEEE Reg. 10 Symp. TENSYMP 2020*, no. June, pp. 1371– 1374, 2020, doi:10.1109/TENSYMP50017.2020.9230464.

[11] M. Rouhani and M. M.Haghighi, “The diagnosis of hepatitis diseases by support vector machines and artificial neural networks,” *2009 Int. Assoc. Comput. Sci. Inf. Technol. Spring Conf. IACSIT-SC 2009*, pp.456–458,2009, doi: 10.1109/IACSIT-SC.2009.25.

[12] V. Vanitha and D. Akila,“Detection and Diagnosis of HepatitisVirus Infection Based on HumanBlood Smear Data in MachineLearning Segmentation Technique,”*2021 9th Int. Conf. Reliab. InfocomTechnol. Optim. (Trends Futur. Dir.ICRITO 2021*, pp. 1–5, 2021, doi:10.1109/ICRITO51393.2021.959648 2.

[13] X. Tian *et al.*, “Using machinelearning algorithms to predict hepatitis B surface antigenseroclearance,” *Comput. Math. Methods Med.*, vol. 2019, 2019, doi: 10.1155/2019/6915850.

[14] K. Polat and S. Güneş, “A hybrid approach to medical decision support systems: Combining featureselection, fuzzy weighted pre- processing and AIRS,” *Comput.Methods Programs Biomed.*, vol. 88, no. 2, pp. 164–174, 2007, doi: 10.1016/j.cmpb.2007.07.013.

[15] M. J. Nayeem, S. Rana, F.Alam, and M.A.Rahman,“Prediction of Hepatitis Disease Using K-Nearest Neighbors, NaiveBayes,Support Vector Machine,Multi-Layer Perceptron and RandomForest,” *2021 Int. Conf. Inf. Commun.Technol. Sustain. Dev. ICICT4SD2021 - Proc.*, pp. 280–284, 2021, doi:10.1109/ICICT4SD50815.2021.9397 013

[16] M. E. Haas *et al.*, “Machine learning enables new insights into genetic contributions to liver fat accumulation,” *Cell Genomics*, vol. 1,no. 3, p. 100066, 2021, doi: 10.1016/j.xgen.2021.100066.

[17] A. A. ABRO, E. TAŞCI, and A. UGUR, “A Stacking-based Ensemble Learning Method for Outlier Detection,” *Balk. J. Electr. Comput. Eng.*, vol. 8, no. 2, pp. 181– 185, 2020,doi:10.17694/bajece.679662.

[18] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[19] L. A. Bull, K. Worden, R. Fuentes, G. Manson, E. J. Cross, and N. Dervilis, “Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data,” J. Sound Vib., vol. 453, pp. 126–150, 2019, doi: 10.1016/j.jsv.2019.03.025.

[20] Yulhendri, Malabay, and Kartini, “Correlated Naïve Bayes Algorithm To Determine Healing Rate Of Hepatitis C Patients,” International Journal of Science, Technology & Management, vol. 4, no. 2, pp. 401–410, Mar. 2023.

[21] Sachdeva, R. K., Bathla., Rani, P., Solanki, V., and Ahuja, R., “A systematic method for diagnosis of hepatitis disease using machine learning,” *Innov Syst Softw Eng*, vol. 19, no. 3, pp. 71–80, Jan. 2023

[22] H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C Virus prediction based on machine learning

[23] Rosidin, S., Muljono, Shidik, G. F., Fanani, A. Z., Zami, F. A., and Purwanto, “Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data,” *International Seminar on Application for Technology of Information and Communication (iSemantic),* Oct. 2021, doi: 10.1109/iSemantic52711.2021.9573196.

[24] Sailasya, G., and Kumari, G. L. A., “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 539–545, 2021.

[25] “UCI Machine LearningRepository: HCV data Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/HCV+data.

[26] Le Cessie S, Van Houwelingen JC (1992) Ridge estimators in logistic regression. J R Stat Soc Ser C Appl Stat 41:191–201

[27] Kumari R, Jose J (2011) Seizure detection in EEG using Biorthogonal wavelet and fuzzy KNN classifier. *Elixir Hum Physiol* 41:5766–5770

[28] Altay O, Ulas M (2018) Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In: 2018 6th International symposium on digital forensic and security (ISDFS). IEEE, pp 1–4

[29] Latha, C.B.C.; Jeeva, S.C. “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”. *Inform. Med.* Unlocked 2019, 16, 100203. [CrossRef]

[30] Ali, R.; Hardie, R.C.; Narayanan, B.N.; De Silva, S. Deep learning ensemble methods for skin lesion analysis towards melanoma detection. In Proceedings of the 2019 IEEE National Aerospace and Electronics Conference

(NAECON), Dayton, OH, USA, 15–19 July 2019; pp. 311–316

[31] Tanuku, S.R.; Kumar, A.A.; Somaraju, S.R.; Dattuluri, R.; Reddy, M.V.K.; Jain, S. Liver Disease Prediction Using Ensemble Technique. In Proceedings of the 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 25–26 March 2022; pp. 1522–1525.

[32] Jongbo, O.A.; Adetunmbi, A.O.; Ogunrinde, R.B.; Badeji-Ajisafe, B. Development of an ensemble approach to chronic kidney disease diagnosis. *Sci. Afr. 2020, 8, e00456*.

[33] Igodan, E.C.; Thompson, A.F.-B.; Obe, O.; Owolafe, O. Erythemato "Squamous Disease Prediction using Ensemble Multi-Feature Selection Approach.: *Int. J. Comput. Sci. Inf. Secur*. IJCSIS 2022, 20, 95–106.

[34] Ashri, S.E.; El-Gayar, M.M.; El-Daydamony, E.M. HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. IEEE Access 2021, 9, 146797–146809.

[35] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023, June). Ensemble learning for disease prediction: A review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI.

[36] T. I. Trishna, S. U. Emon, R. R. Ema,G. I. H. Sajal, S. Kundu, and T. Islam, "Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, pp. 1–7, 2019, doi:10.1109/ICCCNT45670.2019.894445.

[37] M. Rizzetto, S. Hamid, and F. Negro,"The changing context ofhepatitis D," *J. Hepatol.*, vol. 74, no. 5, pp. 1200–1211, 2021, doi: 10.1016/j.jhep.2021.01.014.

[38] M. S. M. Serafim, V. S. dos SantosJúnior, J. C. Gertrudes, V.G.Maltarollo, and K. M. Honorio,"Machine learningtechniques applied to the drug design and discovery of new antivirals: a brief look over the past decade,"