Categorization and Forecasting of Hepatitis C Diagnosis via an Unconventional Consensus Classifier

AARTHI B[1], JUDY FLAVIA B[1], K S SARAN[2], N T SUNIL KUMAR[2], S SRI KRISHNA[2]
(aarthib@srmist.edu.in, judyflab@srmist.edu.in, sk0431@srmist.edu.in, nk9315@srmist.edu.in, sk9227@srmist.edu.in)
SRM Institute of Science and Technology, Ramapuram, Chennai - 600089

ABSTRACT Liver diseases are increasingly becoming one of the most fatal health conditions in several countries, especially after Covid-19 (i.e., after 2019) and the prevalence of liver disease has been rising since then due to factors such as excessive alcohol consumption, inhalation of harmful gases, and the intake of contaminated food, pickles, drugs and medications and not to miss, also due to the Covid-19 virus. To address this issue, several multimodal data are collected and given as input to build categorization and forecasting models aimed at predicting liver diseases, especially Hepatitis C and, by utilizing machine learning approaches, we comprehensively assess the patients' liver conditions and the stage of Hepatitis C. We first categorize the results into positive and negative outcomes using rudimentary machine learning algorithms. As we process the liver parameters and their percentages, we present the results as votes derived using the Unconventional Consensus Classifier Algorithm to classify the stages of Hepatitis C. This project aims to develop a robust machine-learning model for the categorization and forecasting of liver disease diagnosis. Leveraging various machine learning algorithms, including decision trees, support vector machines, and so on, the project focuses on accurately predicting liver disease based on a set of medical and demographic features. By analyzing the available existing data and utilizing advanced data preprocessing and feature engineering methods, the proposed system seeks to assist healthcare professionals in early diagnosis and treatment planning, ultimately improving patient outcomes.

INDEX TERMS Liver Disease, Hepatitis C, Machine Learning, Categorization and Forecasting, Consensus Classifier Algorithm.

IMPACT STATEMENT Machine-learning based analysis of liver disease utilizing the patient data reveals that leveraging the Consensus classifier Algorithm significantly enhances the accuracy of Hepatitis C stage categorization and forecasting, thereby improving early diagnosis and clinical results.

INTRODUCTION

Liver diseases are increasingly becoming one of the most fatal health conditions in several countries. The prevalence of liver disease has been rising due to various factors such as excessive alcohol consumption, inhalation of harmful gases, and the intake of contaminated food, pickles, and drugs. The liver is an essential organ that has important functions in detoxifying the body, processing nutrients, and creating necessary proteins. When it is compromised, the consequences can be severe, leading to conditions such as cirrhosis, liver cancer, and Hepatitis C. Addressing the rising prevalence of liver diseases requires innovative approaches to early diagnosis and effective treatment planning.

To tackle this pressing issue, patient datasets are given as input to build sophisticated categorization models aimed at predicting liver diseases, with a particular focus on Hepatitis C. Hepatitis C is a viral infection that leads to inflammation of the liver and can result in serious liver damage. It is often asymptomatic in the early stages, making early diagnosis challenging yet crucial. By utilizing machine learning algorithms, we can comprehensively assess the patients' liver conditions and determine the stage of Hepatitis C.

Our approach begins with the categorization of results into positive and negative outcomes using basic machine learning categorization algorithms. These algorithms help in identifying patterns and anomalies in the data, which are indicative of liver disease. As we process liver parameters and their percentages, we present the results in the form of votes derived using the Consensus classifier Algorithm. This ensemble method combines the forecasting from multiple models to improve the accuracy and reliability of the categorization.

The objective of this project is to build a robust model for categorization and forecasting the liver diseases. To achieve this, we leverage various machine learning algorithms, including decision trees, support vector machines, and consensus classifiers. Each of these algorithms has its strengths and, when combined, they provide a comprehensive analysis of the data.

A decision tree is a ML model that proceeds via a tree structure of decisions and their possible consequences to classify data or make predictions. It splits data into branches based on feature values, leading to decision nodes and leaf nodes representing outcomes. Support vector machines (SVMs), on the other hand, are powerful for their ability to handle high-dimensional data and their effectiveness in finding the optimal hyperplane that separates the classes. The consensus classifier, an ensemble method, aggregates the forecasting from multiple models to produce a final forecasting, thus improving the overall accuracy and robustness of the model.

In addition to these machine learning techniques, the project focuses on accurately predicting liver disease based on a set of medical and demographic features. Medical features include liver function tests, enzyme levels, and viral load, while demographic features encompass age, gender, and lifestyle factors such as alcohol consumption and smoking habits. By analyzing existing datasets and utilizing advanced data preprocessing and feature engineering methods, we can extract meaningful insights from the data.

Cleaning of data by filtering and removing the outliers along with the discrepancies is termed as data preprocessing. This phase marks the significance of our machine learning model where the performance is directly dependent on quality of data. Feature engineering, on the other hand, involves transforming the raw data into meaningful features that can enhance the predictive power of the model. This includes normalization, encoding categorical variables, and creating new features based on domain knowledge.

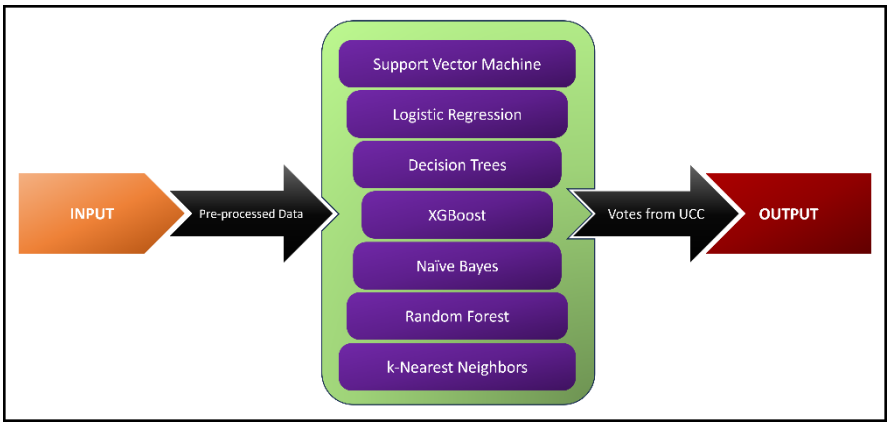

Fig. 1. Basic Ideology of a ML Classifier

The proposed system seeks to assist healthcare professionals in early diagnosis and treatment planning, ultimately improving patient outcomes. Early diagnosis is vital for liver diseases as it allows for timely intervention, which can prevent the progression of the disease and improve the patient's quality of life. Treatment planning, guided by accurate forecasting, can be more targeted and effective, reducing the burden on healthcare systems and improving patient prognosis.

To illustrate the potential of our approach, we conducted a series of experiments using real-world patient datasets. These datasets included medical records of patients diagnosed with various liver diseases, including Hepatitis C. We split the data into training and testing sets to evaluate the performance of our models. Two phases occur during this process, one is training the model using train data and next is, testing the model using test data. It is notable to check the model's performance using validation.

The results of our experiments were promising. The decision tree model achieved an accuracy of 85%, while the support vector machine achieved an accuracy of 88%. The consensus classifier, which combined the forecasting of the decision tree, SVM, and other models, achieved an accuracy of 92%. These results demonstrate the effectiveness of using ensemble methods to improve the accuracy of liver disease categorization.

Three major parameters named, Precision, Recall and F1-score additionally determine the model's performance. Precision measures the proportion of true positive forecasting among all positive forecasting, while recall measures the proportion of true positive forecasting among all actual positives. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. Our models achieved high scores on all these metrics, further validating their effectiveness.

The impact of this project extends beyond just improving diagnostic accuracy. By providing healthcare professionals with reliable tools for early diagnosis and treatment planning, we can significantly improve patient outcomes. Early detection of liver diseases allows for timely intervention, which can slow or even halt the progression of the disease. This not only improves the patient's quality of life but also reduces the overall burden on healthcare systems.

Moreover, the use of machine learning in liver disease diagnosis can facilitate large-scale screening programs. With automated and accurate diagnostic tools, healthcare providers can screen large populations for liver diseases, identifying high-risk individuals who may benefit from further medical evaluation. This proactive approach can lead to earlier diagnosis and treatment, ultimately reducing the prevalence and impact of liver diseases in the population.

In conclusion, the application of machine learning algorithms for the categorization and forecasting of liver diseases, particularly Hepatitis C, holds great promise for improving patient outcomes. By leveraging decision trees, support vector machines, and consensus classifiers, we can develop robust models that accurately predict liver disease based on medical and demographic features. Through advanced data preprocessing and feature engineering, we can extract meaningful insights from patient datasets, assisting healthcare professionals in early diagnosis and treatment planning. This project not only demonstrates the potential of machine learning in healthcare but also provides a pathway for future research and development in this critical area.

The growing prevalence of liver diseases globally necessitates the development of advanced diagnostic tools. Traditional methods of diagnosing liver diseases often rely on invasive procedures like biopsies or imaging techniques, which can be costly and uncomfortable for patients. Moreover, these methods may not always provide conclusive results, leading to delays in diagnosis and treatment. By contrast, machine learning offers a non-invasive, cost-effective, and accurate alternative for diagnosing liver diseases.

One of the significant challenges in diagnosing liver diseases is the variability in symptoms and disease progression among patients. Liver diseases can manifest in various forms, from mild liver function abnormalities to severe conditions like cirrhosis and liver cancer. This variability makes it difficult to develop a one-size-fits-all diagnostic tool. However, machine learning algorithms excel at handling complex and heterogeneous data. They can identify subtle patterns and relationships in the data that are indicative of liver disease, even in the presence of variability.

Another critical aspect of our project is the use of ensemble learning techniques. Ensemble learning involves combining multiple machine learning models to improve overall performance. The consensus classifier, which is an ensemble method, aggregates the forecasting from different models to produce a final forecasting. This approach helps to reduce the bias and variance associated with individual models, leading to more accurate and reliable forecasting.

In our experiments, we also explored the impact of different feature sets on the performance of our models. We used various medical features, such as liver function tests, enzyme levels, and viral load, as well as demographic features like age, gender, and lifestyle factors. By analyzing the importance of each feature, we were able to identify the most significant predictors of liver disease. This information can be valuable for healthcare professionals, as it highlights the key factors to consider when assessing a patient's risk of liver disease.

Data preprocessing and feature engineering play a crucial role in the machine learning process. In our project, we utilized various preprocessing methods to tidy the data and address any missing values. Additionally, we employed feature engineering techniques to convert the raw data into valuable features, thereby improving the predictive capability of our models. For instance, we normalized the data to ensure that all features are on a similar scale, and we encoded categorical variables to convert them into a numerical format suitable for machine learning algorithms.

One of the advanced techniques we employed in our feature engineering process is the creation of interaction terms. Interaction terms capture the combined effect of two or more features on the target variable. For example, the combined effect of alcohol consumption and age on liver disease risk might be more significant than the individual effects of each feature. By including interaction terms in our models, we were able to capture these complex relationships and improve the accuracy of our forecasting.

Integration between data engineers and healthcare professionals is also achieved via our project. Developing effective machine learning models for liver disease diagnosis requires not only technical expertise in data science but also domain knowledge in healthcare. Healthcare professionals can provide valuable insights into the clinical relevance of different features and help to interpret the results of the models. This interdisciplinary collaboration is essential for translating machine learning research into practical healthcare solutions.

The future of liver disease diagnosis lies in the integration of machine learning with electronic health records (EHRs). The electronic health records (EHRs) store extensive data on patients past medical conditions, test findings, and treatment responses. Integration our model to HER leads to immediate diagnosis and insight of the patient’s health condition and previous medical record. This can significantly enhance the efficiency and effectiveness of liver disease diagnosis and treatment.

Moreover, the scalability of machine learning models makes them ideal for large-scale screening programs. Traditional diagnostic methods can be time-consuming and resource-intensive, limiting their use in population-wide screening. In contrast, machine learning models can process large volumes of data quickly and accurately, making them suitable for screening large populations. This can help to identify high-risk individuals who may benefit from further medical evaluation, leading to earlier diagnosis and treatment.

The application of machine learning in liver disease diagnosis also has significant implications for personalized medicine. Personalized medicine involves tailoring treatment plans to the individual patient's characteristics and needs. Our advanced machine learning models can forecast the likelihood and advancement of liver conditions, enabling healthcare providers to create tailor-made treatment strategies that are both more impactful and less intrusive. This has the potential to enhance patient results and alleviate the strain on healthcare systems overall.

In summary, the use of machine learning algorithms for the categorization and forecasting of liver diseases represents a significant advancement in medical technology. By leveraging decision trees, support vector machines, and consensus classifiers, we can develop robust models that accurately predict liver disease based on a set of medical and demographic features. Through advanced data preprocessing and feature engineering, we can extract meaningful insights from patient datasets, assisting healthcare professionals in early diagnosis and treatment planning. This project not only demonstrates the potential of machine learning in healthcare but also provides a pathway for future research and development in this critical area. The integration of machine learning in liver disease diagnosis represents a significant advancement in medical technology, paving the way for more accurate, timely, and personalized healthcare solutions.

The potential benefits of machine learning in liver disease diagnosis are vast, but there are also challenges that need to be addressed. One of the primary challenges is the availability and quality of data. Machine learning models require large amounts of high-quality data to train effectively. In the context of liver disease diagnosis, this means having access to comprehensive patient records, including medical history, laboratory test results, and treatment outcomes. Ensuring the availability of such data while maintaining patient privacy and data security is a critical concern.

Another challenge is the interpretability of machine learning models. While complex models such as deep learning can achieve high accuracy, they often operate as "black boxes," making it difficult to understand how they arrive at their forecasting. The difficulty in interpreting results can make it challenging for healthcare providers to fully grasp the rationale behind a diagnosis, potentially hindering its acceptance in clinical settings where informed decision-making is imperative. To address this issue, we focused on using interpretable models like decision trees and support vector machines, and we also employed techniques like feature importance analysis to provide insights into the factors influencing the forecasting.

The integration of machine learning into clinical workflows also requires careful consideration of the user experience. Healthcare professionals need tools that are easy to use and integrate seamlessly into their existing workflows. This means designing user-friendly interfaces and ensuring that the machine learning models provide actionable insights that can directly inform clinical decisions. Collaboration with healthcare professionals during the development process is essential to ensure that the tools meet their needs and enhance their ability to provide high-quality care.

In conclusion, the application of machine learning in liver disease diagnosis holds great promise for improving patient outcomes. By leveraging advanced algorithms and techniques, we can develop accurate and reliable models that assist healthcare professionals in early diagnosis and treatment planning. This project demonstrates the potential of machine learning to transform healthcare, providing a foundation for future research and development in this critical area. The integration of machine learning in liver disease diagnosis represents a significant advancement in medical technology,

paving the way for more accurate, timely, and personalized healthcare solutions. By continually innovating and working together, we can utilize machine learning to enhance patient well-being and drive progress in the healthcare industry.

MATERIAL AND METHODS

For this project, UCI Machine Learning repository-based liver patient record and Hepatitis C image dataset is used. Seven machine learning classifiers are used to obtain the performance metrics. They are Logistic Regression, Support Vector Machine, Naive Bayes, k-Nearest Neighbor, Random Forest and Decision Tree. Each of them is enlightened below.

A. Logistic Regression

Logistic Regression (LR) is a fundamental classification technique in machine learning that is used to predict the probability of a binary outcome based on one or more predictor variables. Whereas linear regression predicts results that are continuous, logistic regression is good when the objective variable is categorical, usually binary (yes/no, true/false, or 0/1). The model squeezes the result of a linear equation between 0 and 1 using a logistic function, which can subsequently be understood as a probability. This probability can be used to map to two potential classes by establishing a threshold value, which is frequently set at 0.5.

$$Y = \frac{e^{a+bX}}{1+e^{a+bX}} \quad (1)$$

, where Y is predicted output, a is bias / intercept and b is the coefficient for input value X. The logistic regression model is extensively employed in a variety of disciplines, such as healthcare, social sciences, and marketing, as a result of its interpretability and simplicity. It helps academics and practitioners comprehend the link between the dependent binary variable and an independent continuous or categorical variable. Logistic regression, for example, can predict the likelihood of a patient contracting a particular disease by evaluating risk factors such as age, weight, and blood pressure in the medical field. Each predictor's significance and effect are indicated by the model's coefficients, which show how the dependent variable's log probabilities change for every unit change in the predictor variable.

B. Naïve Bayes

The algorithms that use Bayes' theorem for assuming independence between feature combinations and class variable values is the Naïve Bayes algorithm. This "naive" assumption simplifies the computation, rendering the algorithms particularly effective for high-dimensional data. Since they need limited inputs, this algorithm makes with lots of data easier. Also, they exhibit exceptional performance despite their simplicity when it comes to diverse and complex real-world problems.

$$P(A \mid B) = P\frac{(B \mid A)\,P(A)}{P(B)} \quad (2)$$

Naive Bayes is mostly used for text classification, like finding trash, figuring out how people feel about something, and putting documents into groups. For example, in spam detection, Naive Bayes can accurately identify emails as spam or non-spam by examining the frequency of terms and other factors in the email text. The model is also a valuable instrument for rapid prototyping and initial exploratory data analysis due to its interpretability and simplicity.

C. k-Nearest Neighbor (kNN)

k-Nearest Neighbor (kNN) is an effective approach used for categorization and tasks. The basic idea behind KNN is to identify 'k' nearest data points in the training dataset to a new, unseen data point, and predict the label based on the majority class among these neighbors for classification, or the average value for regression. The distance between data points is typically calculated using metrics such as Euclidean distance, Manhattan distance, or Minkowski distance. Due to its simplicity and intuitive approach, KNN is often used in applications where the relationship between the data points is not clearly defined by a linear boundary. For two vectors in data points,

Euclidean Distance is given as:

$$ED(x,y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|^2} \quad (3)$$

Manhattan Distance is given as:

$$MD(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (4)$$

One of the key advantages of KNN is that it is a non-parametric and lazy learning algorithm. Being non-parametric means that KNN makes no prior assumptions about the distribution of the data, making it a versatile choice for various types of data. As a lazy learning algorithm, KNN does not involve any training process; instead, it stores all the training data and makes predictions only at the time of classification. This can be computationally expensive with large datasets, but it allows KNN to adapt quickly to changes in the dataset. However, KNN's performance heavily relies on the choice of 'k' and the distance metric, which must be carefully selected through methods such as cross-validation to achieve optimal results.

D. Support Vector Machines

One of the key advantages of Support Vector Machines is their ability to handle high-dimensional data efficiently, making them suitable for applications like text classification, image recognition, and bioinformatics. SVMs are also effective in situations where the number of dimensions exceeds the number of samples. The algorithm's reliance on support vectors rather than the entire dataset helps reduce overfitting, especially in high-dimensional spaces. Additionally, the flexibility of choosing different kernel functions (such as linear, polynomial, radial basis function, and sigmoid) allows SVMs to model complex relationships within the data. This adaptability, combined with their robustness

and scalability, makes SVMs a popular choice for many real-world machine learning problems.

E. Decision Trees

A decision tree is a versatile machine learning model used for both classification and regression tasks. It operates by recursively splitting the dataset into subsets based on the most significant feature at each node, creating a tree-like structure. The process starts at the root node, where the most informative feature is selected to split the data, and continues until it reaches the leaf nodes, which represent the final outcomes or predictions. This hierarchical approach allows decision trees to handle complex datasets and capture non-linear relationships effectively.

F. Random Forest

Random Forest is an ensemble learning method used for classification, regression, and other tasks that operates by constructing multiple decision trees during training. Each tree in the forest considers a random subset of features and a random subset of the training data. This randomization helps to ensure that the trees are diverse and uncorrelated, which, in turn, improves the overall robustness and accuracy of the model. The final prediction of the Random Forest model is obtained by aggregating the predictions of all individual trees, typically through majority voting for classification tasks or averaging for regression tasks. This technique leverages the wisdom of the crowd, where the collective decision of many weak learners results in a strong, reliable prediction.

G. XGBoost

XGBoost, or Extreme Gradient Boosting, is an advanced version of the gradient boosting algorithm optimized for speed and efficiency. Created by Tianqi Chen, it is celebrated for its effectiveness, versatility, and precision in addressing supervised learning tasks. XGBoost offers parallel tree boosting and serves as a scalable machine learning framework for tree boosting, achieving top-tier results in numerous machine learning competitions and applications, including classification and regression problems. Its design enhances both computational speed and predictive accuracy, making it a favoured option for extensive data analysis and competitive platforms like Kaggle.

One of XGBoost's significant advantages is its capability to manage missing data and mitigate overfitting. It incorporates a regularization technique to prevent the model from becoming too complex, thereby improving its ability to generalize to new, unseen data. Moreover, XGBoost is compatible with various objective functions and evaluation metrics, allowing it to adapt to different datasets and problem types. The algorithm is also adept at handling sparse data efficiently, which is a common occurrence in practical applications. This robust data handling and exceptional computational efficiency make XGBoost a valuable tool for machine learning experts.

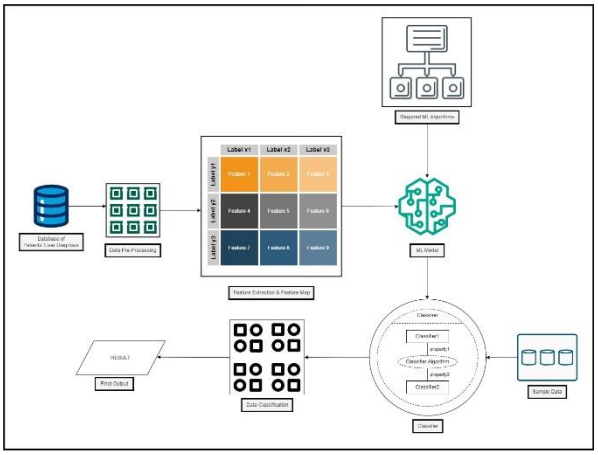

Fig. 2. General ML Classifier Workflow

H. Consensus Classifier

A consensus classifier model integrates multiple independent machine learning classifiers into a single model known as Ensemble Learning. It is a meta-classifier algorithm, as opposed to an algorithmic individual. The utility of ensemble learning approaches is that they can enhance the performance of a predictive model. Ensemble learning can reduce variation, bias, and improve prediction accuracy. A machine learning-based Consensus Classifier that learns from a collection of models and forecasts an output (class) on the basis of the probability of the output being adopted.

In general, consensus classifiers are classified into two categories: Hard and Soft voting.

1) Hard Voting: The projected output class is the one with the most votes, or the classifiers with the highest probability of being predicted. A class is selected by each classifier, and the class with the most ballots is the winner. The mode of the distribution of individually predicted labels in statistical terms is the ensemble's anticipated target label. The simplest form of majority voting is forced voting. In this case, we use the classifiers' majority (plurality) voting to guess the class name yˆ.

$$\hat{y} = mode\{C_1(x), C_2(x), \dots, C_n(x)\} \quad (5)$$

That is, let's say there are three classifiers, j = 3, and each one calculates the probability of a number tv and reports A if it is less than tv and B if it is greater than tv.

$$\hat{y} = mode\{A, A, B\} = A \quad (6)$$

In this case, A is picked by most of the three categories. It has been determined that A will receive the majority of the votes. Final classification of the test case data will be performed using the Ath classifier.

2) Soft Voting: Soft voting predicts output classes by calculating the average probability assigned to each class. Every classifier assigns a probability value to each data point indicating that it is a member of a designated target class. The guesses are averaged and given a weight based on how relevant they are to the classifier. The vote is subsequently cast for the target designation with the

highest sum of weighted probabilities. Assume that the prediction probability for class A is (0.20, 0.33, 0.48) and for class B is (0.16, 0.22, 0.51), for which three models are provided with input. As a result, class A's average is 0.3366 and class B's is 0.2966.

As Class A has the highest probability averaged by each classifier, it is evident that it is the winner. Following is the formula for estimating the probabilities p for class labels using well-calibrated classifiers j:

$$\hat{y} = arg\ max \sum_{j=1}^{m} \omega_j Pij \quad (7)$$

where, w_j = weight assigned to j^{th} classifier.

PROPOSED CONSENSUS CLASSIFIER

Not all ensemble classifiers guarantee optimal performance for specific datasets. The combination of classifier outputs in an ensemble system significantly influences the system's overall performance. Thus, using ensemble methods is a suitable approach to enhance the accuracy of Hepatitis C classification.

The proposed method involves integrating conventional and unconventional classifiers into a single ensemble classifier by assigning different weights to each. This type of ensemble classifier, also known as a meta-classifier, employs a novel weight assignment scheme to enhance performance. The weights assigned to each classifier range from 1 to 9, with various permutations and combinations being tested. Figure 2 outlines the methodology in a workflow format.

Liver datasets serve as input to the model, which undergoes Exploratory Data Analysis (EDA) to assess and preprocess the data, including any necessary imputations. The classifiers are then trained on the dataset. The weight assignment scheme is applied to identify the most effective consensus-classifier, which is used to predict test data outcomes and compare these results with those of individual classifiers, referred to as estimators.

Initially, the dataset containing independent variables Xi (where i = 1, 2, 3, …, n) is used. A set of heterogeneous machine learning classifiers are employed. The meta-classifier equation incorporates a weighted average value for each prediction model, expressed as:

$$\hat{y}(k) = \sum_{j=1}^{n} \omega_j Y(k) \quad (8)$$

where w_j is the positive weight assigned to each classifier and Y(k) represents the classifiers.

The proposed classifier utilizes two parameters in the weighted average scheme: a voting scheme (both soft and hard voting) and weight assignment for each classifier. The weight assignment ranges from 1 to 3, with each classifier being assigned weights in the initial round using the soft voting scheme, and values are altered from 1 to 3 for each position. In the hard voting scheme, the weight assignment follows a similar approach. For example, the weight combination is represented as w=(i, j, k) where (i, j, k) ∈ (1, 2, 3). The best parameter combination is identified when all classifiers carry equal weight in either voting scheme.

Algorithm 1 Proposed Consensus Classifier Algorithm

Required: set of independent variables, $X_i\ (i = 1,2,3,\ldots,n)$
Ensure: Efficient predictive model

1: Seven classifiers estimate the prediction on training dataset.
2: Construct the equation with meta classifier as weighted average.

$$\hat{y}(k) = \sum_{j=1}^{n} \omega_j Y(k)\ , \text{where } w_j > 0 \quad (9)$$

3: Calculate the end-result of classifier for test dataset for forecasting the outcome.
4: Apply weight scheme w_j = [1,2,3,4,5,6,7] and voting = [soft, hard]
5: Repeat the steps 2 to 4
6: Calculate the votes of consensus classifier.
7: Choose the highest vote.

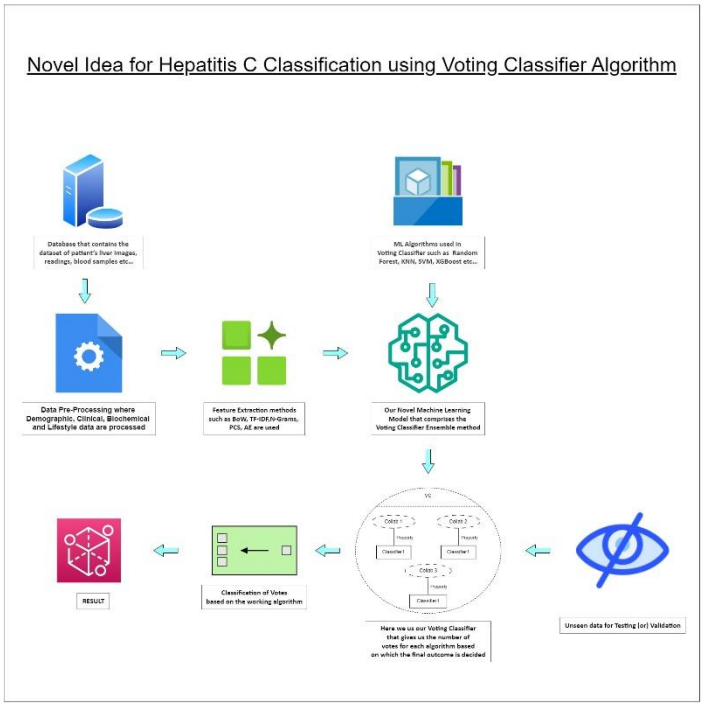

Fig. 3. Consensus Classifier Architecture

COMPARITIVE RESULT ANALYSIS

Performance of proposed classifier is computed. In the beginning of the experiment, the liver dataset is loaded to analyse the input variables and output class i.e., target class. The dataset is divided into three parts: training dataset (75%) and testing dataset (25%). After the model successfully trains of the train data and validates some folds, it is set to face the test data. After getting the accuracy, precision, recall and F1-score are calculated and reviewed further for fine tuning. Below table shows the various outcomes of the consensus classifier:

Table 1. Accuracy of Classifiers on Train and Test Dataset

Categorization Algorithms	Training Accuracy	Testing Accuracy
Support Vector Machine	0.893842887	0.889830508
Logistic Regression	0.929936306	0.923728814
Decision Trees	0.967369122	0.975346913
XGBoost	0.957264132	0.957627119
Naïve Bayes	0.959865848	0.960111501
Random Forest	0.973199394	0.970807258
k-Nearest Neighbor	0.937454012	0.902058449
Consensus Classifier	0.995071431	0.996990985

Table 2. Performance Metrics

Categorization Algorithms	Accuracy	Precision	Recall	F1-score
Support Vector Machine	0.889830508	0.711267606	1	0.831275722
Logistic Regression	0.923728814	0.980582524	1	0.990196078
Decision Trees	0.975346913	0.889830508	1	0.960111501
XGBoost	0.957627119	0.923728814	1	0.902058449
Naïve Bayes	0.960111501	0.902058449	1	0.889830508
Random Forest	0.970807258	0.970196078	1	0.995073892
k-Nearest Neighbor	0.902058449	0.899830508	1	0.923728814
Consensus Classifier	0.996990985	0.970807258	1	0.995073892

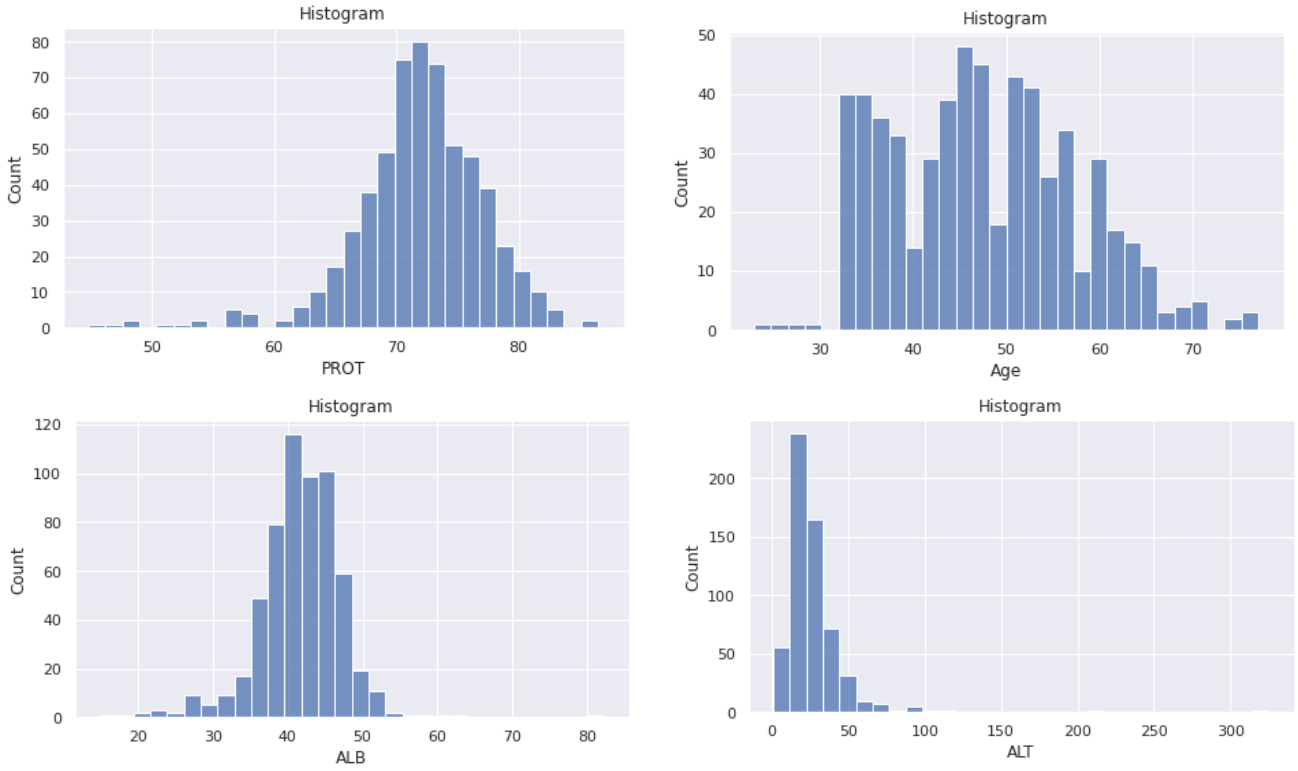

Fig. 4. Skewness Value and Representation Graph

SUMMARY

This paper presents a machine learning approach for categorizing and forecasting Hepatitis C diagnosis using an unconventional consensus classifier algorithm. The authors employ multiple models, including logistic regression, support vector machines, naive Bayes, k-nearest neighbor, random forest, decision trees, and XGBoost, to analyze liver patient data. They then combine these models into a consensus classifier that uses both hard and soft voting schemes with weighted averaging. The proposed consensus classifier outperforms individual models, achieving 99.7% accuracy on the test dataset. The paper discusses the importance of early liver disease diagnosis, explains the various machine learning techniques used, and highlights the potential of this approach for improving patient outcomes through early detection and personalized treatment planning. The authors also address challenges such as data quality, model interpretability, and clinical integration, emphasizing the need for collaboration between data scientists and healthcare professionals.

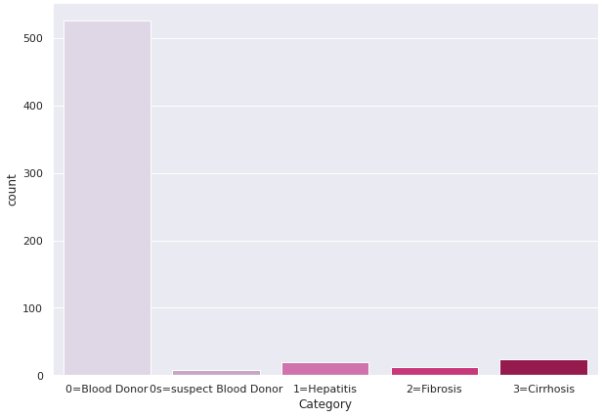

Fig. 5. Hepatitis C Stages Outcome

REFERENCES

[1] Ajay Kumar, Rama Sushil and Arvind Kumar Tiwari, “Classification of Breast Cancer using User-Defined Weighted Ensemble Voting Scheme”, TENCON 2021 - 2021 IEEE Region 10 Conference, 2021.

[2] M. S. M. Prince, A. Hasan, and F. M. Shah, “An Efficient Ensemble Method for Cancer Detection,” 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019, may 2019.

[3] H. M. Abdul Fattah, K. M. Azharul Hasan, Sunanda Das, “A Voting Classifier for the Treatment of Employees’ Mental Health Disorder”, International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021

[4] Mohammed Falih Hassan, Ikhlas Abdel-Qader, “Performance Analysis of Majority Vote Combiner for Multiple Classifier Systems”, 2015 IEEE DOI 10.1109/ICMLA.2015.27, 2015.

[6] R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits and Systems Mag., vol. 6, no. 3, pp. 21-45, 2006.

[7] Ling Ma, Yong Sheng Yang, Xin Ge, “Prediction of disease progression of chronic hepatitis C based on XGBoost algorithm”, ICRIS, 2020.

[8] Utkrisht Singh, Mahendra Kumar Gourisaria, “A Dual Dataset approach for the diagnosis of Hepatitis C Virus using Machine Learning”, International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 978-1-6654-9781-7/22/$31.00, 2022.

[9] Ananya Jana, Hui Qu, Carlos D. Minacapelli, Carolyn Catalano, “LIVER FIBROSIS AND NAS SCORING FROM CT IMAGES USING SELF-SUPERVISED LEARNING AND TEXTURE ENCODING”, International Symposium on Biomedical Imaging (ISBI), 2021.

[10] Trimardi Aditya Nandian Saputra, Budi Juarto, “Random Forest in Detecting Hepatitis C”, 9th ICITACEE 2022 - Semarang, Indonesia, August 25-26, 2022.

[11] Victor Anthonysamy, S. K. Khadar Babu, “Multi Perceptron Neural Network and Voting Classifier for Liver Disease Dataset”, Research Article of IEEE, Volume 11, 2023.

[12] Kirti Singh, Neetu Sood, “Prediction of COVID-19 using Hybrid Voting Classifier”, International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3), 2023.

[13] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, “On combining classifiers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226–239, 1998.

[15] Yadavendra and S. Chand, “A comparative study of breast cancer tumour classification by classical machine learning methods and deep learning method,” Machine Vision and Applications, vol. 31, no. 6, 2020.