Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 1 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i01.51383

PAPER

Smart Diagnosis: Leveraging Machine Learning for Early Detection of Hepatitis in Healthcare

Hemang Mehta[1], Vyom Shah[1], Sashikala Mishra[2], Nirmal Swain[3], Chinmay Kulkarni[4], Debabrata Swain[1](✉)

[1]Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

[2]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India

[3]Department of Information Technology, Vardhaman College of Engineering, Telengana, Hyderabad, India

[4]Department of Information Technology, University of Cumberlands, Williamsburg, KY, USA

debabrata.swain7@yahoo.com

ABSTRACT

There is high variance related to the detection and prevention of human diseases. This is a concerning factor considering hepatitis. It is a disease that affects the functioning of the liver, which causes the deaths of millions of people in a year, around the world. Conventional methods are slow and inaccurate to a large extent. Machine learning (ML) is a process where a machine is trained using large amounts of data for it to predict about the disease. This paper aims at developing a hybrid machine-learning model using a stacking-based classifier. The hepatitis dataset is available in the UCI ML repository used in this work. A few data pre-processing steps were implemented on the dataset to create an optimum database. This includes imputing missing values, balancing the dataset using the synthetic minority over-sampling technique (SMOTE) and scaling the dataset using robust scaler. For selecting the optimum features, Chi-Square and Pearson Correlation tests were performed. The proposed model has reported a classification accuracy of 98.7%.

KEYWORDS

support vector machine (SVM), logistic regression (LR), nearest neighbour classifier, hepatitis

1 INTRODUCTION

The liver is a major metabolic organ of the human body [1]. Its synthesises proteins and biochemicals that are necessary for human organ growth. At present there are no efficient long-term alternatives for human liver. Short-term alternatives may include liver dialysis, which is costly and infeasible to maintain for a long time. In case of liver failure, the last viable option left is liver transplant. Therefore, livers are irreplaceable and should be kept as far from harm as possible. The worldwide mortality rate per year due to liver disease is approximately two million. Every year people suffer from various liver-related diseases. One of the common types of liver syndrome found in all age groups in the society is hepatitis. Hepatitis is a common condition leading to liver inflammation. It results from various infectious and non-infectious factors, leading to a variety of health issues, some of which are

Mehta, H., Shah, V., Mishra, S., Swain, N., Kulkarni, C., Swain, D. (2025). Smart Diagnosis: Leveraging Machine Learning for Early Detection of Hepatitis in Healthcare. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(1), pp. 26–40. https://doi.org/10.3991/ijoe.v21i01.51383

Article submitted 2024-07-25. Revision uploaded 2024-10-20. Final acceptance 2024-10-20.

life-threatening. Hepatitis can affect a person in two ways. The acute way, where a recent infection presents itself with a rapid onset, or the chronic way, where the long-lasting asymptomatic condition progresses to a life-threatening hepatic disease. The hepatitis virus comprises five primary strains known as A, B, C, D, and E. These strains differ significantly in aspects such as transmission methods, illness severity, geographical prevalence, and prevention strategies. The primary types of hepatitis are A, B, and C that all the people around the world.

Hepatitis A is caused by personal contact, ingestion of the virus, or consumption of contaminated food. The risk is higher for children and people who are travelling or living in places with high infection rates. The symptoms are generally mild sickness. hepatitis B is transmitted when the bodily fluids of an infected person enter our body [2]. People with multiple sexual partners and those who inject drugs are at a higher risk. The symptoms can range from sickness and infections to liver damage or failure. The possible cause of hepatitis C is due to blood transfer from a sick person. The risk is particularly high for people who recklessly inject drugs. The symptoms can be severe, including long-term infections, liver failure, or even cancer [3].

Estimates indicate that around 296 million lives were lost due to hepatitis B in 2019, with the addition of 1.5 million more each year. The death rate has reached an estimate of 820,000 a year despite vaccines being developed. There were approximately 58 million cases of hepatitis C and 1.5 million new cases each year since 2019. The death rate has reached an estimate of 290,000 a year [4]. There is no effective vaccine for hepatitis C as of yet. Detecting hepatitis as early as possible is of utmost importance. It can be done so with various tests. These tests are done with the purpose of looking for disease before a person develops symptoms, diagnosing the root cause, and determining the right treatment required. A few types of tests are blood tests, liver ultrasounds, liver biopsies, etc. In these blood tests, doctors measure 'HAV antibodies' for Hepatitis A, 'antibodies, antigens, etc. of HBV' for hepatitis B, and 'antibodies and genetic material of HCV' for hepatitis C [5].

The medical decision made by a physician based on the above set of tests is not always accurate. It depends upon the experience of the physician and is time-consuming. Due to unavailability of symptoms, it is not possible to detect the presence of hepatitis with patients in time. This delay in diagnosis mostly results in a large number of casualties.

Nowadays, by combining different health-related parameters, it is possible to predict the occurrence of any disease using machine learning (ML) algorithms. ML is a branch of artificial intelligence that mainly focuses on creating intelligence within a computer by identifying the useful patterns from a complex dataset.

In this study, different ML-based models are applied for accurate identification of hepatitis disease. The objective of this study is to create an ML-based healthcare monitoring system to assist the doctors during the diagnosis of hepatitis disease with more accuracy by consuming less time.

2 LITERATURE REVIEW

Different ML algorithms are applied for the identification of hepatitis C. A genetic algorithm was used to select 11 optimal features out of 44. For model tuning, the GridSearchCV algorithm was used. The data balancing issue was not addressed in this work [6].

A system is created for quantifying and detecting hepatic fibrosis that has been implemented using a support vector machine (SVM) model. The classifier was verified through 10-fold cross-validation to summarise the performance. However, this study has not processed the data, such as scaling the data, balancing the data, and feature extraction [7].

An ML-based disease identification system for Hepatitis C was developed. The model was tested at various stages, each containing varying subsets of data from the complete dataset, and was cross-validated five times to gain the average performance of these models. Yet, this study considers numerous features that dampen the accuracy of the model by adding unnecessary information [8].

An ML-based liver illness identification system was implemented. This system was developed using four models, wherein the highest accuracy was achieved by the random forest classifier (RFC). The vastness of the dataset and use of excessive features affected the accuracy of the model [9].

An ML-based model for the classification of liver disease was used. The highest accuracy in both these instances is achieved using k-nearest neighbours (KNN) and RFC, respectively. However, despite the vastness of data, due to the small size of instances for each class in multiclass labels, the accuracy was heavily affected. So is true for the surplus of features due to the independency of each feature with respect to each other [10].

Different ML algorithms were applied for predicting the presence of different hepatitis syndromes. Cross-validation is performed to find the generic behaviour of the model. Random forest has shown the highest accuracy of 98% [11].

Bagging and boosting-based classifiers were used for predicting the status of hepatitis B disease. The AUC score was calculated for finding the efficiency of the applied models. The highest accuracy reported by the decision tree (DT) classifier [12].

A hepatitis diagnosis system using ML and deep learning algorithms was implemented. The highest accuracy of 87% was found with the ANN model [13].

Custom models for the early prediction of hepatitis disease were implemented. The missing values were removed from the dataset, correlated features were identified, and the dataset was balanced during the pre-processing phase. Finally, the XGBoost and neural network were optimised using a hyperparameter tuning algorithm [14].

A cascaded ML model was developed for hepatitis virus diagnosis. The Synthetic Minority Over-sampling Technique (SMOTE) method was used for data balancing. Finally, the proposed RF-MLR model was formed using the ABC algorithm [15].

Multiple ML algorithms were applied for predicting hepatitis disease. For dataset balancing, the SMOTE algorithm was applied. The highest accuracy was shown by the logistic regression (LR) algorithm [16].

Forward feature selection was employed to choose important features from the dataset. After that, different ML algorithms were applied to classify the hepatitis patients and healthy persons [17].

After performing a detailed investigation of the above discussed literature papers, the following research gaps are identified.

1. In most of the research works, feature selection was not performed.
2. In a few of the research works, the data balancing was not performed, which results in a biased classifier.
3. In most of the work, hyperparameter tuning was not performed.

3 PROPOSED SYSTEM

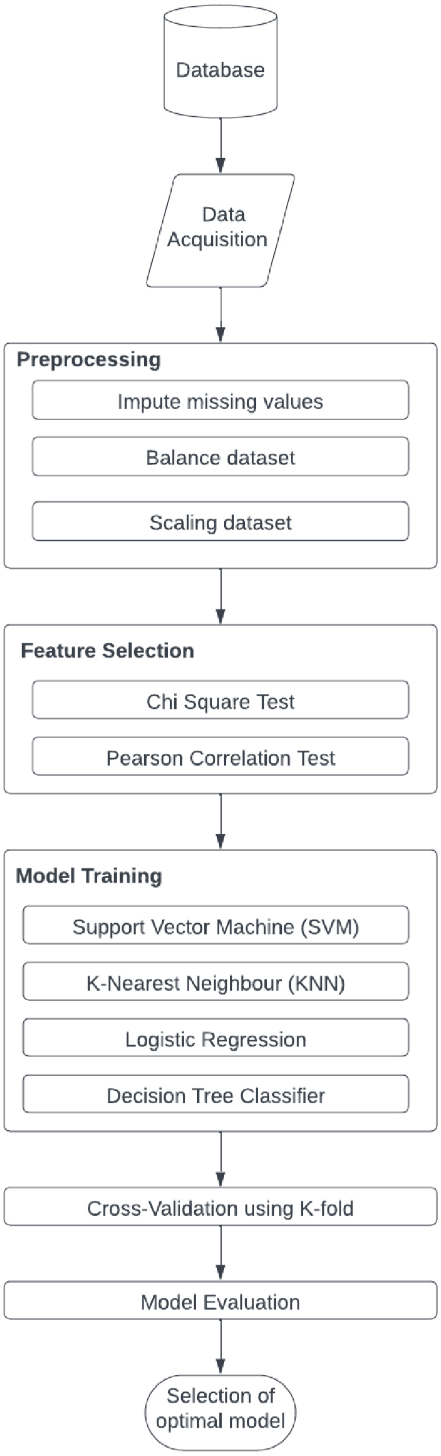

Fig. 1. System architecture of proposed work

The different phases of the proposed system architecture are shown in the Figure 1.

3.1 Data acquisition

The dataset for the proposed research was collected from the UCI ML repository, a widely recognised source for obtaining reliable and resourceful data [18]. The diverse set of features provides an all-inclusive summary of the health status of blood donors and Hepatitis C patients across different diagnostic categories, including 'Blood Donor,' 'Suspect Blood Donor,' 'Hepatitis,' 'Fibrosis,' and 'Cirrhosis.' This dataset includes 615 records, each characterised by 14 features, of which 2 are categorical, representing patient diagnosis ('Category') and gender ('Sex'), while the remaining 12 features are numerical. The numerical attributes include patient demographic information such as age, as well as laboratory values reflecting various blood parameters that are given in Table 1.

Table 1. Laboratory measured features

Feature	Description
ALB	Albumin levels
ALP	Alkaline phosphatase levels
ALT	Alanine aminotransferase levels
AST	Aspartate aminotransferase levels
BIL	Bilirubin levels
CHE	Cholinesterase levels
CHOL	Cholesterol levels
CREA	Creatinine levels
GGT	Gamma-glutamyl levels
PROT	Total protein levels

3.2 Data acquisition

In this phase, the missing values are handled, the dataset is balanced and scaled, and finally useful features are identified.

i) **Handling missing values:** The dataset attained from the UCI repository underwent a series of pre-processing steps to address specific challenges and enhance its suitability for subsequent analysis. Initially, 31 missing values were found across multiple features. To fix this, the missing values x_i in these features were replaced by x_{mean} (mean of the remaining values) within each corresponding feature. By doing this, the impact of missing data on further analyses was reduced, and a thorough and correct representation of the dataset was guaranteed. The categorical features (category, sex) were converted to a binary format of zero and one. This encoding step contributed to creating a standardised and numerically compatible dataset, setting the stage for subsequent model training and evaluation. Figure 2 represents the count of missing values found in the dataset.

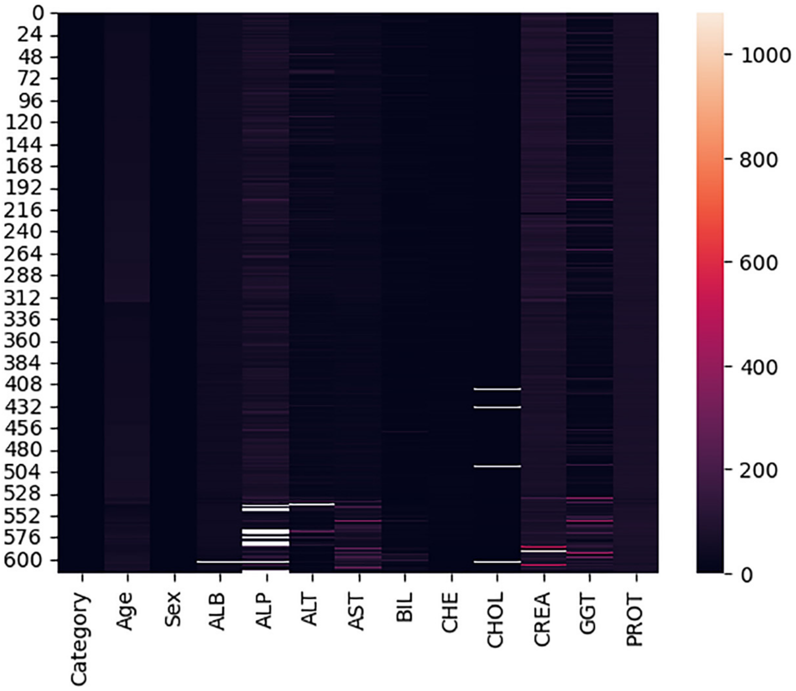

Fig. 2. Missing values visualisation

ii) **Data balancing:** A significant issue identified in the dataset was the notable class imbalance, wherein 540 records corresponded to patients who did not have hepatitis and 75 records to patients who did, as shown in Figure 3. The ratio of healthy to diseased patients is 36:5. Bias may be introduced during model training and assessment as a result of this imbalance. A SMOTE was used to correct this [19]. The SMOTE generates artificial instances by altering the vector between the chosen data point and its neighbours, choosing random examples from the minority class, and determining the KNN. To provide a fairer distribution and lessen the possibility of biased model results, this oversampling strategy attempted to balance the representation of both groups. After putting these pre-processing procedures into practice, an improved dataset of 1080 records were produced as shown in Figure 4.

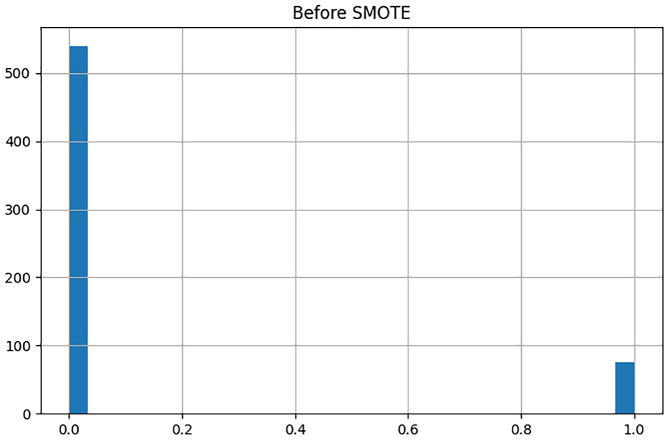

Fig. 3. Balancing using the SMOTE function

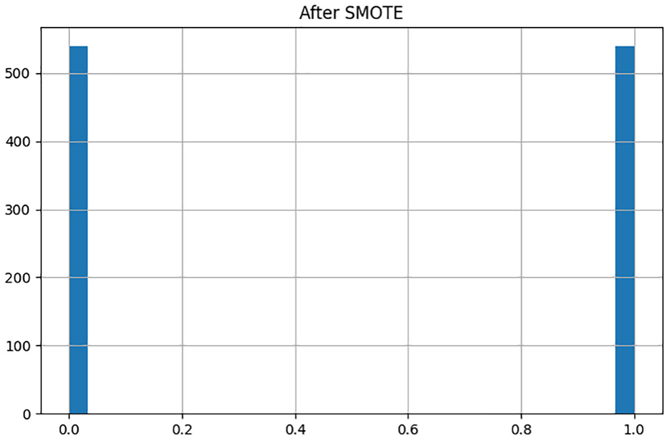

Fig. 4. After balancing using the SMOTE function

iii) **Data scaling:** To efficiently handle the outliers, the robust scaling method is employed here [20]. There are outliers detected in different features as shown in Figure 5. To efficiently handle these values, robust scaling method is applied. This scaling technique is advantageous in the presence of outliers as it utilises statistics, making the features less affected by extreme values. It scales the data (x) based on the interquartile range (IQR), calculated as the difference between the 3rd (X_{75}) and 1st (X_{25}) quartile. The formula for the robust scaler is stated in equation (1).

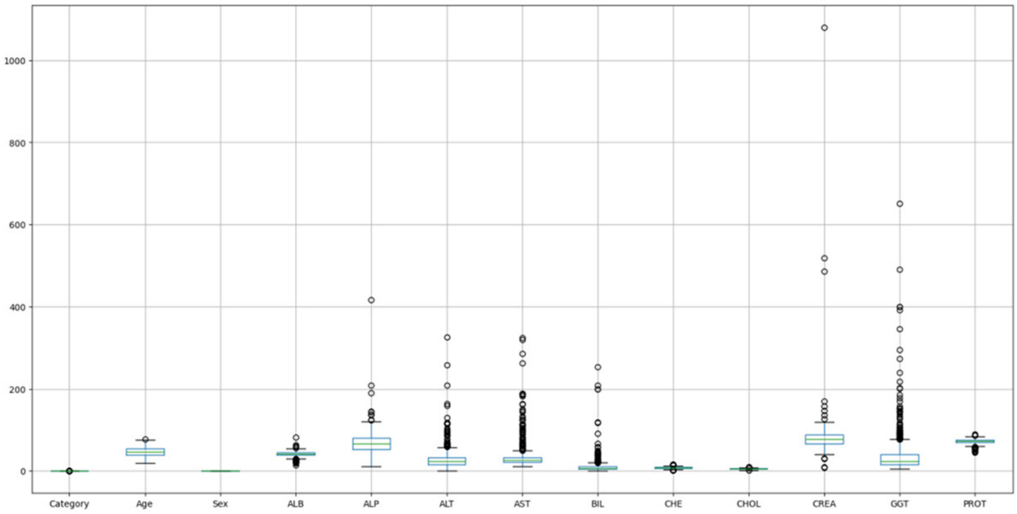

Fig. 5. Distribution of values in each feature

$$X_{sc} = \frac{X - Q_1(X)}{Q_3(X) - Q_1(X)} \tag{1}$$

Where X_{sc} is the scaled value, X is the original value, $Q_1(X)$ is the 1st quartile, and $Q_3(X)$ is the 3rd quartile.

iv) **Feature selection:** Inessential features, when present while training the model, unknowingly impact the accuracy. Two methods used for feature selection are as follows:

a. Chi-squared test: The chi-squared test is employed to ascertain the existence of a relationship between two categorical features, where one typically serves as the target feature [21]. In the dataset, the categorical feature 'Sex' was considered, with 'Category' being the target feature. The test is grounded in hypothesis testing, wherein the null hypothesis (H0) posits that the two categorical features are independent. The chi-squared value (x_f^2) is calculated from the following equation (2).

$$x_f^2 = \sum \frac{(O - E)^2}{E} \tag{2}$$

Where, O represents observed values, E denotes expected values, and f is the degree of freedom.

A greater difference between observed and expected values yields a higher x_f^2 value, supporting the acceptance of the null hypothesis, indicating independence between the selected and target features.

The chi-square value achieved for the features 'Sex' and 'Category' was $2.66*10^{-5}$. Therefore, the null hypothesis was rejected, and the feature 'Sex' had been removed.

b. Pearson correlation: Correlation measures the similarity between two continuous variables. Pearson correlation coefficient (PCC) measures the association amount between different features. This method is used for identifying redundant or highly correlated features, as their presence can adversely affect model performance [22].

The Pearson correlation coefficient is calculated using the formula given in equation (3):

$$PCC = \frac{\sum(a - \bar{a})(b_i - \bar{b})}{\sqrt{\sum(a_i - \bar{a})^2 \sum(b - \bar{b})^2}} \tag{3}$$

Here, a_i and b_i represent individual data points for the two features being compared, $\bar{a}$ and $\bar{b}$ are their means. The numerator computes the covariance between the two features, and the denominator normalises the result using the product of their standard deviations. The resulting PCC value ranges between –1 and 1. Where –1 stands for a strong negative relationship, 1 denotes a strong positive relationship, and 0 signifies no relation. If two features have a correlation value greater than the threshold value, it indicates that they are highly correlated and one of the features can be removed without affecting the resulting accuracy. If the correlation value is close to 0, that means that the two features are not dependent on each other. After applying the correlation function and displaying the results, we get the following map as shown in Figure 6.

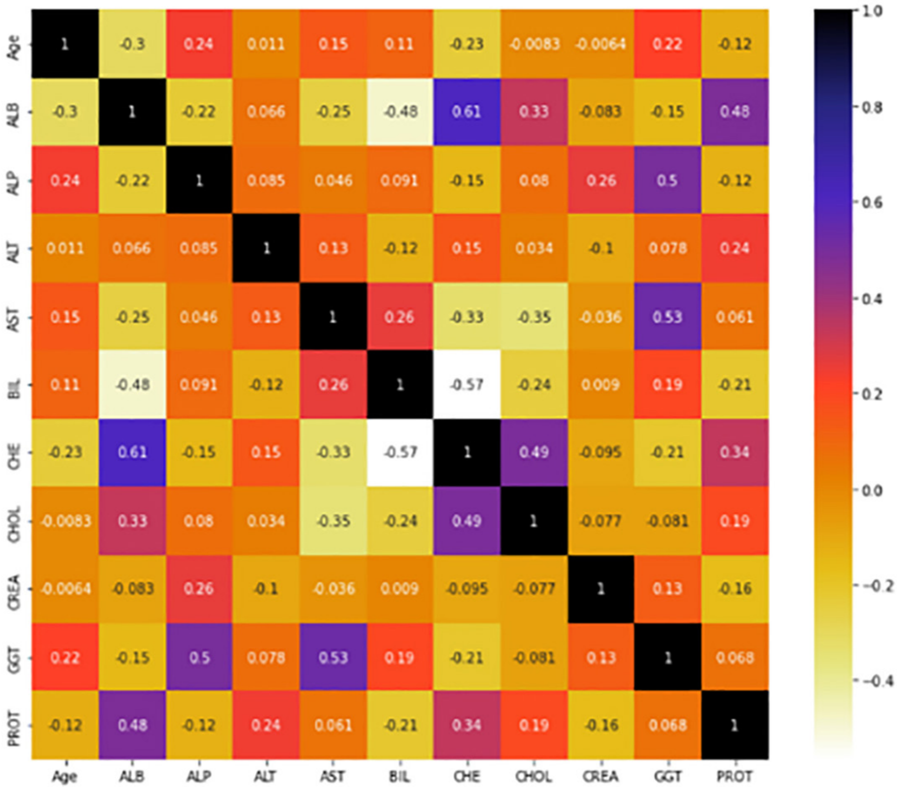

Fig. 6. Pearson correlation matrix

The final feature selection decision is made based on the threshold value of 0.5, which is demonstrated in the equation (4).

$$(fi) = \{fi \in F \mid fcorr \geq Threshold\} \tag{4}$$

fi = individual feature
F = set of all selected features
fcorr = correlation score of a feature

In the given map, we can see three features, {'BIL', 'CHE', 'GGT'} have high correlation with other features and hence are eligible to be removed. The final features that are considered for model training are {'ALB', 'ALP', 'ALT', 'AST', 'CHOL', 'CREA', 'PROT'} based on chi-square score and Carl-Pearson correlation score.

3.3 Data acquisition

i) Support vector machine: SVM is an ML algorithm commonly applied for classification tasks [23]. It identifies an optimal hyperplane to segregate data points into distinct classes in an N-dimensional space. The three essentials hyperparameters of this algorithm are 'C,' 'kernel,' and 'gamma.' The trade-off between smoothness of the decision border and classification accuracy is controlled by the 'C' parameter. Points that have a greater 'gamma' value and are near the decision border are given higher priority. The type of decision boundary is determined by the 'kernel' parameter, which is essential to the algorithm's functioning. The degree of similarity between two points is computed using the following function. Let us take two points, x_1 and x_2, then the function will be as given in equation (5).

$$\mathrm{k}(\mathrm{x}_1, \mathrm{x}_2) = e^{\left(-\frac{\|\mathrm{x}_1 - \mathrm{x}_2\|^2}{2\sigma^2}\right)} \tag{5}$$

Where σ is variance and $\|\mathrm{x}_1 - \mathrm{x}_2\|$ is the Euclidean distance between the two points. The optimal parameters, identified through a grid search, are {C: 10, gamma: 0.1, kernel: 'rbf'}, employing the radial basis function ('rbf'). KNN: KNN is an ML-based classification algorithm [24]. Fundamentally, KNN finds the nearest neighbours having a majority to classify the given data point. For calculating the distance with the neighbours, the algorithm uses the Euclidean distance formula as shown in equation 6. Three hyperparameters were selected: "leaf_size," "n_neighbors," and "p" in order to fine-tune the model.

To arrange the points from the training dataset for effective neighbour searches, KNN builds a KD-Tree data structure. The number of points in a leaf of the KD-Tree is defined by the 'leaf_size' option. A deeper KD-Tree with more nodes and better query performance is produced by smaller "leaf_size" values, whereas bigger "leaf_size" values result in a shallower tree with fewer nodes and lower memory use. 'n_neighbors' represents the number of neighbouring points.

$$\mathrm{D}(\mathrm{g}, \mathrm{h}) = \sqrt{\Sigma(g_i - h_i)^2} \tag{6}$$

The best parameters chosen are {'leaf_size': 1, 'n_neighbors': 2, 'p': 1}.

ii) Logistic regression: LR is a probabilistic based classification algorithm [25]. It mainly classifies a data point into two classes using the sigmoid functions shown in equation 7.

$$\sigma_x = \frac{1}{1 + e^{-x}} \tag{7}$$

Here, σ_x is the sigmoid value, and *x* is the data point.

The crucial hyperparameters of the LR algorithm are "penalty," "C," and "solver." 'C' represents the regularisation parameter, 'penalty' represents the type of regularisation and 'solver' is the optimisation term.

The optimum values chosen by GridSearchCV are {'C': 1, 'penalty': 'l1', 'solver': 'linear'}.

iii) Decision tree: DT classifier is a non-linear classification algorithm [26]. It mainly does the classification decision using different features present at different nodes of the tree. The inclusion of any feature into the tree can be decided by calculating its information gain using the Gini index score as shown in equation 8.

$$G(f) = 1 - \sum_{i=1}^{k} (p_i)^2 \tag{8}$$

Where G(f) = Gini index of feature X; k = number of output classes. p_i = probability score of any data point to be classified under class 'i'.

4 RESULTS AND DISCUSSION

To find the efficiency of the proposed system, different factors such as accuracy, precision, recall, and F1 score are calculated. Accuracy measures the correctness of the entire classification [27]. Precision measures the correctness of identifying positive instances out of all positive predictions done by the model. Recall is the measure of accurate positive case identification out of all actual positive cases [28]. The mathematical expressions for accuracy, precision, and recall are mentioned in equations 9, 10, and 11.

$$Accuracy = (p + s)/(p + q + r + s) \tag{9}$$

$$Precision = p/(p + q) \tag{10}$$

$$Recall = p/(p + r) \tag{11}$$

The different observations regarding the model performance are given in the Tables (2–4).

Table 2. Accuracy after balancing the dataset at a 70–30 split

	Accuracy	F1 Score	Precision Score	Recall Score	Cohen Kappa
SVM	93.510	0.930	0.946	0.915	0.869
KNN	97.530	0.974	0.962	0.987	0.950
LOG	95.980	0.957	0.960	0.954	0.919
DT	95.987	0.957	0.954	0.961	0.919

Table 3. Accuracy after feature selection from the dataset at a 70–30 split

	Accuracy	F1 Score	Precision Score	Recall Score	Cohen Kappa
SVM	95.987	0.960	0.987	0.934	0.919
KNN	98.148	0.982	0.970	0.994	0.962
LOG	95.061	0.951	0.963	0.940	0.901
DT	97.530	0.975	0.987	0.964	0.950

Table 4. Accuracy after hyperparameter tuning the models at a 70–30 split

	Accuracy	F1 Score	Precision Score	Recall Score	Cohen Kappa
	Before Tuning	After Tuning			
SVM	97.530	99.382	0.956	0.970	0.942
KNN	99.382	99.074	0.956	0.970	0.942
LOG	95.061	95.370	0.956	0.964	0.948
DT	97.83	95.62	0.968	0.944	0.913

The ROC curve, a plot of the TP rate against the FP rate at various threshold values, is another tool used to assess the performance of the model.

Figures 7 to 10 represent the ROC curves of the models before performing hyperparameter tuning. Figures 11 to 14 represent the ROC curves of the models after performing the hyperparameter tuning.

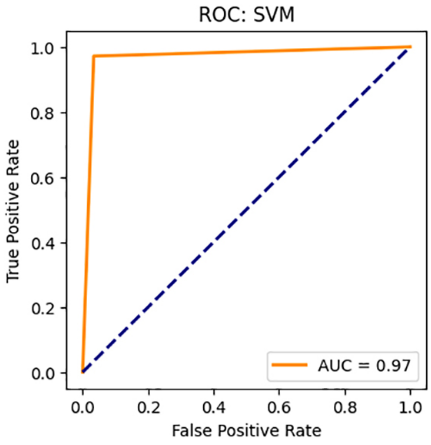

Fig. 7. ROC of SVM

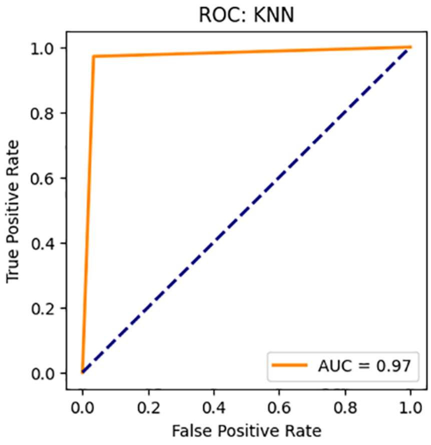

Fig. 8. ROC of KNN

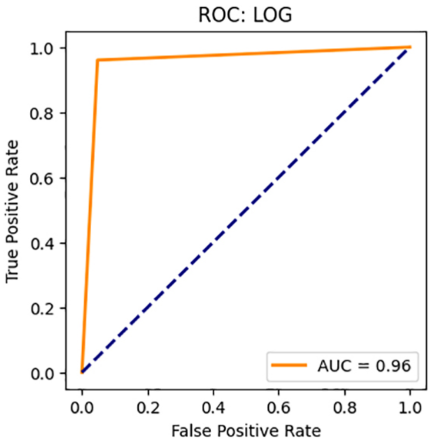

Fig. 9. ROC of LR

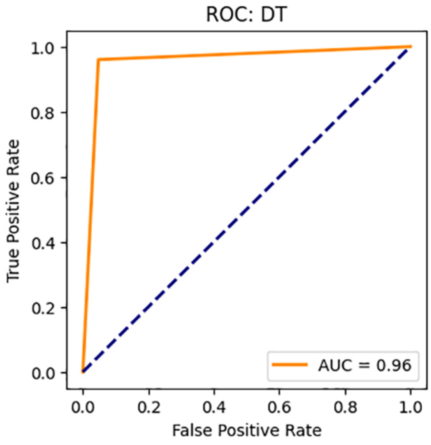

Fig. 10. ROC of DT

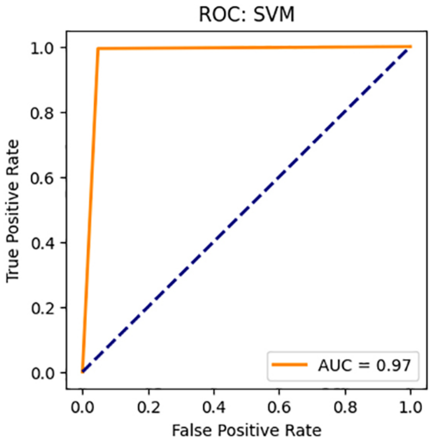

Fig. 11. ROC of SVM

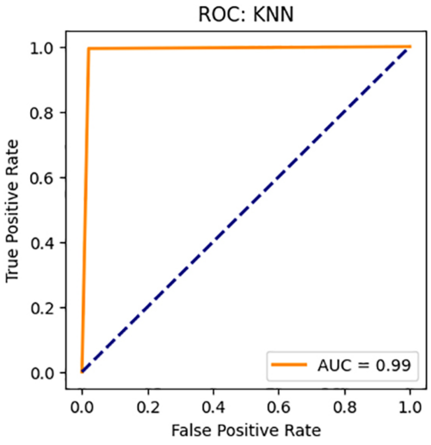

Fig. 12. ROC of KNN

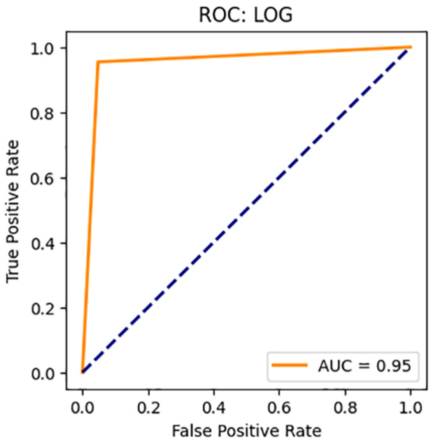

Fig. 13. ROC of LR

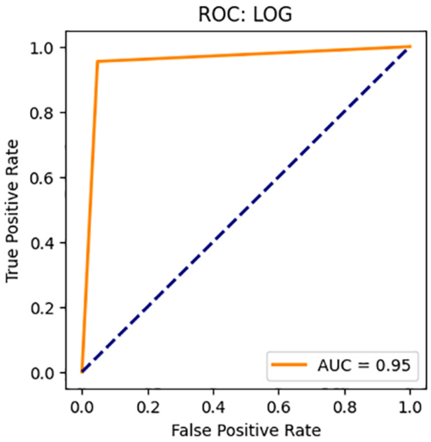

Fig. 14. ROC of DT

5 COMPARATIVE ANALYSIS

After analysing the performance of the proposed hybrid classifier, the comparative analysis of the result is given in Table 5. Even though the different tuned ML models are used for classification, the accuracy was not promising. Data imbalance and a greater number of feature considerations are the primary reasons behind this degradation of performance [6]. The lack of feature selection in the SVM model has limited its capacity to attain greater accuracy rates, even with the use of 10-fold cross-validation for performance assessment. This constraint indicates that feature selection strategies should be included in order to improve model performance [7]. Inadequate data scaling and balancing is the reason why this model is unreliable, which shows why data pre-processing is an important step [8]. While applying several ML classifiers, the lack of data preparation and balancing results in superior performance in predicting liver fibrosis. It also produced skewed data and inaccurate forecasts [9]. Due to the train-test data ratio of 50:50, the model's capacity was hindered despite using ensemble-based learning approaches. This shows why suitable data-splitting techniques are important for both model training and evaluation [10].

The comparative analysis of the performance of the different discussed methods and the proposed method is shown in Table 5.

Table 5. Comparative analysis of performance of discussed methods

Reference	Methodology	Accuracy	Size of Dataset
[6]	ML models with GridSearchCV	86.05%	1385 records
[7]	SVM using 10-fold cross validation	95.4% AUROC	987 records
[8]	Machine learning model for diagnosing the stage of liver fibrosis in patients with chronic viral hepatitis C	80.56%	689 records
[9]	Detection of Liver Fibrosis using DTC, RFC, LRC and SVC Classifiers	91.1%	920 records
[10]	Binary and Multi Class Classification using different ensemble-based learning.	50.83%	1385 records
	Proposed Methodology	**99.382%**	**1230 records**

6 CONCLUSION

In this paper, a comprehensive analysis of the dataset was conducted where the necessary data pre-processing procedures were carried out to address the CKD dataset's shortcomings, which included missing values, imbalanced data, and the existence of outliers. The hyperparameters were then adjusted to produce an efficient KNN model. After applying feature selection, the seven most important features were selected and were fed to the models. The model has a 99.382% prediction accuracy, which can be used in real life by medical assistants and doctors, reducing the disease from spreading further and worsening for the patients.

7 REFERENCES

[1] World Health Organization, "Hepatitis," 2024. [Online]. Available: https://www.who.int/health-topics/hepatitis#tab=tab_1. [Accessed: 15-Oct-2024].

[2] Mayo Clinic, "Hepatitis B – diagnosis and treatment," 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/hepatitis-b/diagnosis-treatment/drc-20366821. [Accessed: 15-Oct-2024].

[3] Testing.com, "Hepatitis testing," 2024. [Online]. Available: https://www.testing.com/hepatitis-testing/. [Accessed: 15-Oct-2024].

[4] World Health Organization, "Hepatitis B," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b. [Accessed: 15-Oct-2024].

[5] S. O. Olatunji *et al.*, "Preemptive diagnosis of Hepatitis C using machine learning techniques: A retrospective study in Saudi Arabia," in *2023 International Conference on Smart Computing and Application (ICSCA)*, Hail, Saudi Arabia, 2023, pp. 1–6. https://doi.org/10.1109/ICSCA57840.2023.10087834

[6] S. Gawrieh *et al.*, "Automated quantification and architectural pattern detection of hepatic fibrosis in NAFLD," *Annals of Diagnostic Pathology*, vol. 47, p. 151518, 2020. https://doi.org/10.1016/j.anndiagpath.2020.151518

[7] V. Tsvetkov, I. Tokin, and D. Lioznov, "Machine learning model for diagnosing the stage of liver fibrosis in patients with chronic viral hepatitis C," *Preprints*, pp. 1–12, 2021. https://doi.org/10.20944/preprints202102.0488.v1

[8] N. Li *et al.*, "Machine learning assessment for severity of liver fibrosis for chronic HBV based on physical layer with serum markers," *IEEE Access*, vol. 7, pp. 124351–124365, 2019. https://doi.org/10.1109/ACCESS.2019.2923688

[9] S. C. Nandipati, C. XinYing, and K. Khai Wah, "Hepatitis C Virus (HCV) prediction by machine learning techniques," *Applications of Modelling and Simulation*, vol. 4, pp. 89–100, 2020. http://arqiipubl.com/ojs/index.php/AMS_Journal/article/view/122

[10] 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," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT*), Kanpur, India, 2019, pp. 1–7. https://doi.org/10.1109/ICCCNT45670.2019.8944455

[11] N. Ali, D. Srivastava, A. Tiwari, A. Pandey, A. K. Pandey, and A. Sahu, "Predicting life expectancy of hepatitis B patients using machine learning," in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE*), Ballari, India, 2022, pp. 1–4. https://doi.org/10.1109/ICDCECE53908.2022.9793025

[12] S. S. Nigatu, P. C. R. Alla, R. N. Ravikumar, K. Mishra, G. Komala, and G. R. Chami, “A comparitive study on liver disease prediction using supervised learning algorithms with hyperparameter tuning,” in *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT*), Gharuan, India, 2023, pp. 353–357. https://doi.org/10.1109/InCACCT57535.2023.10141830

[13] L. Chen, P. Ji, and Y. Ma, “Machine learning model for hepatitis C diagnosis customized to each patient,” *IEEE Access*, vol. 10, pp. 106655–106672, 2022. https://doi.org/10.1109/ACCESS.2022.3210347

[14] T.-H. S. Li, H.-J. Chiu, and P.-H. Kuo, “Hepatitis C virus detection model by using random forest, logistic-regression and ABC algorithm,” *IEEE Access*, vol. 10, pp. 91045–91058, 2022. https://doi.org/10.1109/ACCESS.2022.3202295

[15] R. K. Sachdeva *et al.*, “A systematic method for diagnosis of hepatitis disease using machine learning,” *Innovations Syst. Softw. Eng.*, vol. 19, pp. 71–80, 2023. https://doi.org/10.1007/s11334-022-00509-8

[16] H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, “Hepatitis C Virus prediction based on machine learning framework: A real-world case study in Egypt,” *Knowl. Inf. Syst.*, vol. 65, pp. 2595–2617, 2023. https://doi.org/10.1007/s10115-023-01851-4

[17] Fedesoriano, “Hepatitis C prediction dataset,” kaggle, 2020. [Online] Available: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

[18] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, “An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6651–6672, 2023. https://doi.org/10.1109/TKDE.2022.3179381

[19] T. Sarkar, M. Roozbehani, and M. A. Dahleh, “Robustness scaling in large networks,” in *2016 American Control Conference (ACC)*, Boston, MA, USA, 2016, pp. 197–202. https://doi.org/10.1109/ACC.2016.7524915

[20] D. Mihalko, “Chi-square tests-of-fit for location-scale families using type-II censored data,” in *IEEE Transactions on Reliability*, vol. 42, no. 1, pp. 76–80, 1993. https://doi.org/10.1109/24.210274

[21] X. Zhi, S. Yuexin, M. Jin, Z. Lujie, and D. Zijian, “Research on the Pearson correlation coefficient evaluation method of analog signal in the process of unit peak load regulation,” in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Yangzhou, China, 2017, pp. 522–527. https://doi.org/10.1109/ICEMI.2017.8265997

[22] Q. Wang, “Support vector machine algorithm in machine learning,” in *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, 2022, pp. 750–756. https://doi.org/10.1109/ICAICA54878.2022.9844516

[23] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019, pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747

[24] X. Zou, Y. Hu, Z. Tian, and K. Shen, “Logistic regression model optimization and case analysis,” in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, 2019, pp. 135–139. https://doi.org/10.1109/ICCSNT47585.2019.8962457

[25] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, “Overview of use of decision tree algorithms in machine learning,” in *2011 IEEE Control and System Graduate Research Colloquium*, Shah Alam, Malaysia, 2011, pp. 37–42. https://doi.org/10.1109/ICSGRC.2011.5991826

[26] S. Kumar, N. Neware, A. Jain, D. Swain, and P. Singh, "Automatic helmet detection in real-time and surveillance video," in *Machine Learning and Information Processing, Advances in Intelligent Systems and Computing*, D. Swain, P. Pattnaik, and P. Gupta, Eds., Springer, Singapore, 2020, vol. 1101, pp. 51–60. https://doi.org/10.1007/978-981-15-1884-3_5

[27] D. Swain, S. S. Bijawe, P. P. Akolkar, A. Shinde, and M. V. Mahajani, "Diabetic retinopathy using image processing and deep learning," *International Journal of Computing Science and Mathematics (IJCSM)*, vol. 14, no. 4, pp. 397–409, 2021. https://doi.org/10.1504/IJCSM.2021.120686

[28] D. Swain *et al*., "Cardiovascular disease prediction using various machine learning algorithms," *Journal of Computer Science*, vol. 18, no. 10, pp. 993–1004, 2022. https://doi.org/10.3844/jcssp.2022.993.1004

8 AUTHORS

Hemang Mehta is currently continuing his Bachelors in Technology in Computer Science and Engineering from Pandit Deendayal Energy University, Gandhinagar. His area of interest is machine learning and computer vision.

Vyom Shah is currently continuing his Bachelors in Technology in Computer Science and Engineering from Pandit Deendayal Energy University, Gandhinagar. His area of interest is machine learning and computer vision.

Sashikala Mishra is working as a Professor at the Computer Science department at Symbiosis Institute of Technology, Pune. Her areas of interest are machine learning and computer vision.

Nirmal Swain is working as an Assistant Professor at the Department of Information Technology, Vardhaman College of Engineering, Telengana. His areas of interest are machine learning and deep learning.

Chinmay Kulkarni is a PhD student at Department of Information Technology, University of Cumberlands, Williamsburg, Kentucky. His areas of interest are machine learning and computer vision.

Debabrata Swain is working as an Assistant Professor at Pandit Deendayal Energy University, Gandhinagar. His area of interest includes machine learning and deep learning (E-mail: debabrata.swain7@yahoo.com).