Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Hepatitis Diagnosis: A Comprehensive Review of Machine Learning Classification Algorithms

Hiveen Saleem Sadiq[1], Adnan Mohsin Abdulazeez[2]
hayveen.sadiq@gmail.com, adnan.mohsin@dpu.edu.krd
[1]IT department, Technical College of Duhok, Duhok Polytechnic University, Duhok, Iraq
[2]Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq

Article Information

Submitted : 3 May 2024
Reviewed : 10 May 2024
Accepted : 15 Jun 2024

Keywords

Hepatitis, Machine Learning, Classification, Algorithms

Abstract

Hepatitis is a liver-related medical disorder caused by inflammation, often caused by hepatitis virus infection or an unknown source. There are five primary hepatitis viruses: A, B, C, D, and E. Machine learning (ML) algorithms have emerged as a promising tool for hepatitis diagnosis, leveraging vast datasets and complex patterns. This review examines the application of ML classification algorithms in hepatitis diagnosis, focusing on challenges faced in traditional diagnostic approaches and the potential of ML techniques to address these. Various ML algorithms, including decision trees, support vector machines, neural networks, Naïve Bayes, random forest, K-nearest neighbor, and logistic regression and ensemble methods, are analyzed for their efficacy in hepatitis classification tasks. Key considerations such as data preprocessing, feature selection, and performance evaluation are also discussed. The review aims to provide clinicians, researchers, and healthcare stakeholders with a comprehensive understanding of ML algorithms' role in hepatitis diagnosis and improving patient outcomes.

A. Introduction

Hepatitis also referred to as inflammation of the liver, is a potentially self-limiting condition that may progress to cirrhosis, fibrosis, or liver cancer. Hepatitis viruses are the most prevalent cause of hepatitis worldwide, but autoimmune diseases and noxious substances are also capable of inducing the disease[1]. Viruses A, B, C, D, and E are the five primary varieties of hepatitis viruses. These are the most worrisome due to the burden of disease and mortality, as well as the potential for outbreaks and the dissemination of visitors. Specifically, in a vast number of individuals, types B and C are responsible for chronic illness and are also the leading causes of cancer and liver cirrhosis. Hepatitis A and hepatitis E are often caused by the consumption of contaminated water and food, whereas Hepatitis B, Hepatitis C, and Hepatitis D are caused by coming into touch with infected bodily fluids via direct entry into the bloodstream [2]. The main means of transmission for these viruses include blood contamination during medical procedures, contamination of equipment, transfer from parent to child during childbirth, and transmission between family members, particularly from adults to children. Additionally, sexual contact may also be a common source of infection [3].

The infection might manifest with little or absent symptoms, but may also present with symptoms such as abdominal discomfort, black urine, profound exhaustion, jaundice, nausea, or vomiting. Hepatitis may arise from several factors, such as excessive alcohol use, negative reactions to drugs, and bacterial or viral infections. Early detection of hepatitis greatly enhances the likelihood of a good recovery [4]. Utilizing machine learning for the diagnosis of hepatitis illness will enhance efficiency and aid novice physicians. Human mistake may occur due to factors such as fatigue, lack of expertise, and distraction. Identifying the hepatitis virus continues to be a difficult task for inexperienced medical professionals. The study presents a method for diagnosing hepatitis illness by using machine learning methods [5]. Machine Learning (ML) is a technique that enables a system to autonomously learn by identifying patterns and correlations within the data via the use of various algorithms [6]. This would facilitate the automated diagnosis of any illnesses, with careful consideration given to the selection of parameters and the tool used for assessing these data [7].

This research investigates three distinct techniques used for predicting Hepatitis: Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM). Technology can simplify health problems by creating tools like artificial intelligence and expert systems. These systems can think like experts, helping doctors diagnose diseases like liver disease. K-nearest neighbors (KNN) is a powerful method for categorizing data items based on their closest neighbors[8]. When diagnosing hepatitis, KNN is a classification method used for diagnosing hepatitis by comparing individuals to existing instances [9]. It helps identify the condition and quantify risk, but its performance can be influenced by distance metric selection and neighboring number. Naive Bayes, a classifier using Bayes' theorem, is effective for text classification and can be used to determine the likelihood of a patient being classified into a specific subtype. However, its assumption of feature independence may not always be valid in complex medical datasets, potentially limiting its effectiveness[10]. The Support Vector Machine (SVM) is a powerful technique used in binary and multiclass classification problems, aiming to identify the ideal hyperplane to separate different classes in the feature space. It is widely used in medical fields, especially in diagnosing hepatitis due to its ability to handle large-dimensional data and make decisions using kernel functions[11]. SVM's resistance to overfitting and generalization make it appealing for clinical decision support systems. However, performance depends on kernel function and regularization parameters. KNN, Naive Bayes, and SVM have unique benefits and compromises. [10]. Although KNN offers a simple classification method, while Naive Bayes offers probabilistic reasoning and efficiency for large-dimensional data. Support Vector Machines excel in identifying decision limits and extrapolating to unfamiliar data. Healthcare practitioners and researchers can make informed decisions about these algorithms for hepatitis diagnosis and contribute to machine learning-driven healthcare solutions[12].

This paper aims to review existing literature on hepatitis diagnosis. The rest of this paper is organized to describe the background theory of machine learning and the algorithms utilized for detecting hepatitis[13]. Following this, in section 2, various research employing Support Vector Machines, K-nearest Neighbors, Random Forest, and Naive Bayes for hepatitis diagnosis are outlined and compared in Table 1. Finally, section 3 describes a discussion session, a conclusion with limitations, and a list of references.

B. Research Method

Machine learning classification algorithms are crucial in the field of hepatitis diagnosis as they assist medical practitioners in accurately and efficiently assessing patients. These algorithms use several computers approaches to analyses patient data and forecast the probability of hepatitis occurrence or severity [14]. Widely used algorithms in this context are logistic regression, which calculates the likelihood of hepatitis based on patient characteristics; decision trees, which divide the data into separate categories for classification; and support vector machines, which determine the best hyperplane for separating classes. In addition, ensemble approaches like as random forests and gradient boosting machines are used to merge numerous models in order to improve prediction accuracy[15]. These techniques, including k-nearest neighbors, naive Bayes, and artificial neural networks, provide various methods for diagnosing hepatitis. Each algorithm has its own distinct advantages and is suitable for different types of data. By using these machine learning methods, medical practitioners may accelerate the process of diagnosing, enhance the effectiveness of treatment plans, and ultimately enhance the results for patients in managing hepatitis [16].

Hepatitis diagnosis is being done by researchers using machine learning classification algorithms as datasets and algorithms are required to learn and extract comments for analysis [17]. The diagnosis of hepatitis is a crucial medical undertaking that often requires precise and prompt determination of the type and severity of the illness [18]. Machine learning classification algorithms have become significant tools for supporting healthcare practitioners in recent years. These algorithms use patterns and connections in patient data to forecast hepatitis kinds, stages, and consequences. Logistic Regression, a basic approach in machine learning, is often used for diagnosing hepatitis [19]. It is especially beneficial for jobs involving binary classification, such as differentiating between diseased and noninfected persons. Logistic regression may assist doctors in decision-making processes by using input data such as liver enzyme levels and viral markers to model the link with the probability of hepatitis present and offer probabilistic predictions.

Decision Trees provide a straightforward framework for diagnosing hepatitis, since they divide the feature space according to different clinical factors[20]. Every node in the decision tree corresponds to a certain characteristic, while the branches indicate the many potential outcomes or options [21]. Clinicians may analyze the series of choices that result in a diagnosis by navigating through the tree. Decision trees are particularly advantageous due to their transparency and capacity to handle both numerical and categorical data.

Support Vector Machines (SVMs) are very effective in diagnosing hepatitis by identifying the most ideal hyperplane that can accurately differentiate patients who are infected from those who are not infected in the feature space. Support Vector Machines (SVMs) are very efficient in datasets with a large number of dimensions, since they are capable of detecting intricate decision boundaries[22]. SVMs improve the resilience of hepatitis classification models by maximizing the separation between different classes[23]. This allows for accurate predictions even when dealing with noisy or overlapping data.

The Random Forest algorithm, which is a kind of ensemble learning approach, enhances the accuracy of hepatitis detection by combining numerous decision trees [24]. Random Forest improves generalization performance by combining the predictions of multiple trees, hence reducing overfitting. This method is especially advantageous when working with extensive datasets that include a wide range of patient groups and intricate relationships between different characteristics. The Random Forest method, a kind of ensemble learning technique, improves the accuracy of hepatitis diagnosis by aggregating many decision trees. Random Forest enhances the generalization capability by amalgamating the predictions of several trees, hence mitigating overfitting [25]. This approach is particularly beneficial when dealing with large datasets that include diverse patient groups and complex interrelationships among various attributes. Machine learning algorithms are supervised and unsupervised, with each algorithm's reliability demonstrated in Figure 1.

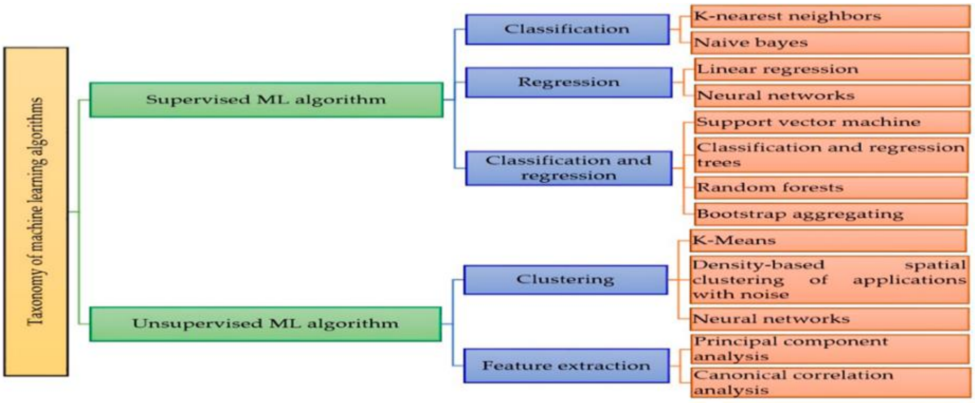
Taxonomy of machine learning algorithms
Supervised ML algorithm
Classification
K-nearest neighbors
Naive bayes
Regression
Linear regression
Neural networks
Classification and regression
Support vector machine
Classification and regression trees
Random forests
Bootstrap aggregating
Unsupervised ML algorithm
Clustering
K-Means
Density-based spatial clustering of applications with noise
Neural networks
Feature extraction
Principal component analysis
Canonical correlation analysis

Figure1. Classification of machine learning techniques [26]

Support Vector Machine (SVM) for Hepatitis Diagnosis

Support vector machines (SVM) are fundamentally linear classifiers or Linear Learning Machines (LLM). In Support Vector Machines (SVM), a hyperplane is selected to separate two classes in a way that minimizes the chance of misclassification. This is achieved by maximizing the functional gap between the two classes. The training data points that lie on the margins of this ideal hyperplane are referred to as support vectors[27]. The learning process involves identifying the support vectors. SVM uses Kernel functions to translate the input vector from the input space to a higher dimensional feature space, especially for non-linearly separable data. The precise selection of the Kernel function is a crucial step in the effective construction of a Support Vector Machine (SVM) for a certain classification assignment [28].

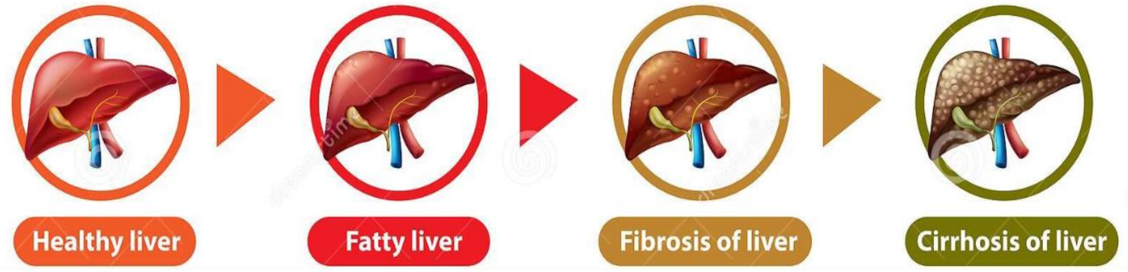

Figure 2. Stage of Liver Disease leading to cirrhosis of Support Vector Machine (SVM)[29]

K-nearest neighbors (KNN) for Hepatitis Diagnosis

K-Nearest Neighbors (KNN) This classifier does categorization using a three-step process. During the first stage, the system calculates the K-value. During step 2, the algorithm calculates the distance between each test sample and all the training data. It then ranks the distances. In step 3, the class name for the test sample is determined using the majority vote technique. The K-nearest neighbors (KNN) algorithm is a flexible machine learning technique that may be used for diagnosing hepatitis. KNN utilizes patient data, including clinical traits and laboratory findings, to classify people by comparing their characteristics to those of known instances in the dataset [30]. The KNN method uses a predetermined number of closest neighbors to a data point and assigns the most common class label. Its success in detecting hepatitis is determined by rigorous training and assessment methods. The interpretability of the KNN model aids physicians in diagnosing hepatitis, facilitating its integration into clinical practice. Continuous improvement ensures its relevance across patient groups, improving healthcare decision-making and patient care [31, 32].

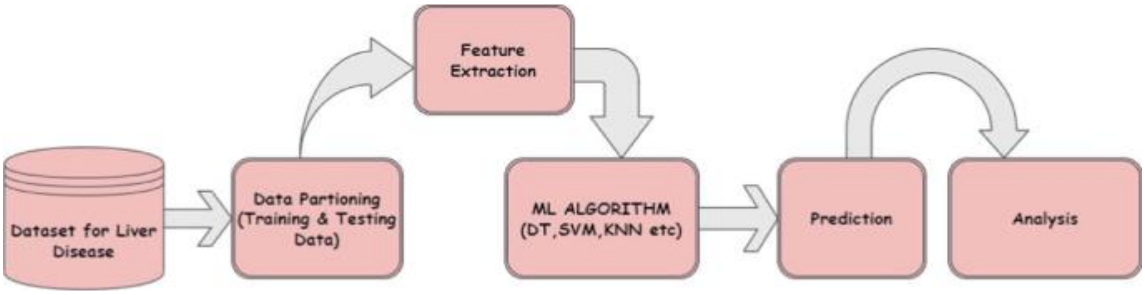
Dataset for Liver Disease
Data Partioning (Training & Testing Data)
Feature Extraction
ML ALGORITHM (DT,SVM,KNN etc)
Prediction
Analysis

Figure 3. Detection of Liver Disease Using of K-nearest neighbors (KNN)[33]

Random forest (RF) is a versatile and user-friendly machine learning technique that generates. Most of the time, a terrific outcome can be achieved even without hyperparameter adjustment. Due to its simplicity and versatility, random forest is also one of the most often used algorithms. The random forest is an ensemble method that combines many decision trees, with each tree being built using a different set of randomly selected variables. A decision tree is a flowchart that has a vector-like form. The vector x = $(x1, x2, \ldots, xp)$ T represents the predictor variables in a n-dimensional space, whereas y represents the real-valued answer [25]. Figure 2 is an illustration of a random forest.

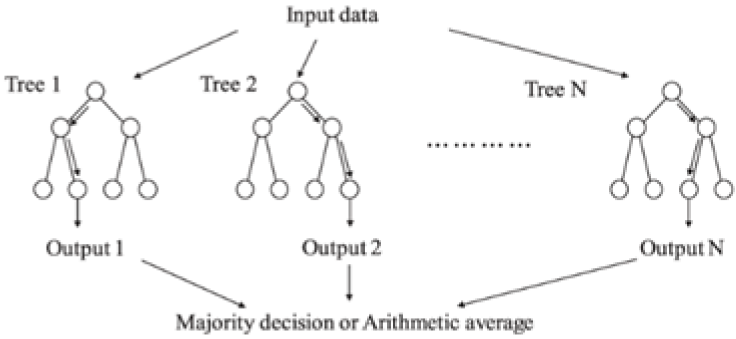

Figure 4. Illustration of random forest [34]

The strong ensemble learning approach Random Forest (RF) may help diagnose hepatitis. In medical diagnostics, Random Forest trains numerous decision trees and combines their predictions to improve accuracy and resilience. Each decision tree is trained on a subset of data and random characteristics to reduce overfitting and improve generalization. This ensemble technique lets Random Forest handle complicated clinical datasets with many characteristics and interactions. Random Forest provides a consensus conclusion that is typically more trustworthy than any single decision tree by aggregating tree forecasts. Random Forest may increase hepatitis diagnostic accuracy and predictability, enabling earlier identification, intervention, and treatment. Random Forest's capacity to uncover key diagnostic traits helps understand disease causes and provide personalized treatment choices, improving hepatitis patient outcomes[35, 36].

Naive Bayes (NB) for Hepatitis Diagnosis

Naive Bayes classifier is a probabilistic model based on Bayes' Theorem, simplifying learning by assuming class-independent features. Although often considered a bad assumption, it is practical and efficient in supervised learning. It uses maximum likelihood for parameter estimation, allowing for non-Bayesian techniques.[37]. Naive Bayes classifiers, despite their seemingly oversimplified assumptions, perform better than expected in complex real-world scenarios. Recent research shows their theoretical underpinnings, and they can diagnose hepatitis using Bayes' theorem to estimate a patient's hepatitis category based on clinical and laboratory findings [38]. Naive Bayes is a powerful diagnostic paradigm for hepatitis, offering a simple, interpret able approach based on observable characteristics. Its computing efficiency and scalability make it ideal for real-time or high-throughput diagnostics, helping doctors identify and stratify patients early. [39].

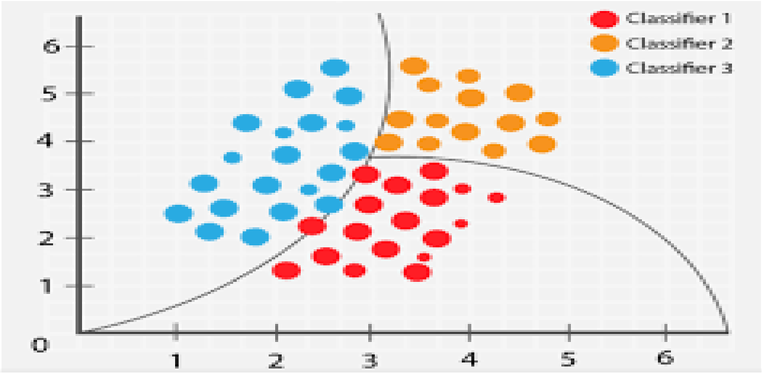

Figure5. Illustration of Naive Bayes (NB)[40]

The most comprehensive publicly accessible datasets for hepatitis diagnosis were used for the comparison against several baselines and state-of-the-art.

In [41] the authors developed a prediction framework for Hepatitis C Virus among healthcare professionals in Egypt using machine learning methods. The data was gathered from the National Liver Institute at Menoufiya University, Egypt, and included 859 patients and 12 characteristics. The model was assessed using Naïve Bayes, random forest (RF), K-nearest neighbor, and logistic regression. The results showed that the Sequential Forward Selection (SFS) feature selection method resulted in greater accuracy, with the RF classifier reaching an accuracy of 94.06% with a 0.54 second learning time.

In [42] researchers improved predictive models by eliminating missing values, using ranker search and info-gain feature selection, and applying classification techniques like K-Nearest Neighbors, Naive Bayes, Multi-Layer Perceptron, and Random Forest. The model's performance was evaluated for accuracy, precision, recall, F1-score, and ROC. The Random Forest algorithm demonstrated superior performance, achieving a classification accuracy of 92.41%.

In [43] the authors used machine learning to make predictions based on historical data since there was so much of it. Results from evaluations have varied, according on the literature reviewed. In order to optimize performance, a strategy is suggested and put into action using training data and important model variables. The Naïve Bayes approach has an area under the curve (AUC) of 72.5%, according to calculations and analysis, whereas the k-nearest neighbor (KNN) strategy had an AUC of 63.19%.

In [44] the authors used the UCI library's hepatitis C dataset for research, involving four stages: data preparation, data mining, assessment, and system design. Data balancing was used at the preprocessing stage, followed by three data mining techniques: multi-layer perceptron, Bayesian network, and decision tree. The 10-fold cross-validation approach was used for evaluation. The MATLAB programming language was used for user interface development. The over-sampling technique improved data mining algorithms' performance in illness prediction. The random forest model achieved an impressive 99.9% accuracy in the O-dataset, with a perfect specificity rate of 100%.

In [45] the researchers used a dataset of 113 observations and 5 variables to analyze hepatitis test results using support vector machines (SVM) and random forest algorithms. The SVM model, using a gaussian radial basis function kernel, achieved an accuracy rate of 99.55%. The linear kernel SVM achieved an accuracy of 99.13%, while the random forest algorithm achieved an accuracy of 98.43%. The SVM model with a polynomial kernel exhibited the least accurate performance, achieving an accuracy rate of 96.64%. Both SVM and random forest algorithms were used for prediction.

In [46] the researchers aimed to enhance the accuracy of naïve Bayes and KNN algorithms by analyzing 155 public data from the UCI Repository, including 19 attributes like age, gender, and more. The experiments used a confusion matrix to determine accuracy, precision, and recall. The Naïve Bayes algorithm achieved an accuracy of 74.19% and an average error of 25.81% higher than the K-Nearest Neighbor algorithm, which had an average value of 54.84% and an average error of 45.18%. The K-Nearest Neighbor algorithm improved accuracy and average errors from previous studies.

In [47] the researchers presented a classification platform for hepatitis using random forest, decision tree, and support vector machine algorithms. It reveals a strong correlation between certain characteristics and hepatitis diagnosis, suggesting potential improvements in early-stage diagnosis. Random forest showed the highest accuracy across various feature sets, with accuracies of 96.1% and 95.7%. The use of different feature sets for each model showed significant performance enhancements while maintaining consistency, indicating potential for improved accuracy with minimal loss using fewer features.

In [48] the researchers used the UCI dataset to evaluate machine learning methods for predicting hepatitis. They used support vector machine, logistic regression, K-nearest neighbor, and random forest. The study found significant improvements in classifier efficiency with class balancing, with logistic regression with SMOTE demonstrating the highest accuracy at 93.18%.

In [49] the researchers used suggested a comparative analysis was conducted to examine different machine learning tools and neural networks, as suggested. The mean square error and the accuracy rate comprised the performance metric. As instruments for classification and prediction in the diagnosis of Hepatitis disease, Machine Learning (ML) algorithms including Support Vector Machines (SVM), K

Nearest Neighbor (KNN), and Artificial Neural Network (ANN) were taken into consideration. Based on the results obtained, it can be concluded that among all the models evaluated, ANN exhibits the highest level of accuracy and performance. It provides a 96 percent prediction accuracy and a minimum mean square error.

In [50] the authors compared the sensitivity, specificity, and accuracy of Decision Tree and K-Nearest Neighbor classifiers in detecting Novel Hepatitis C. Data was obtained from Kaggle and analyzed using MATLAB and a standard dataset. Results showed significant differences in accuracy, sensitivity, and specificity between the two classifiers. K-NN demonstrated superior performance in accuracy, sensitivity, and specificity, while Decision Tree showed significant differences in sensitivity and specificity. The study concluded that K-NN is a more accurate and effective method for detecting novel hepatitis C.

In [51] the authors aimed to evaluate the accuracy of machine learning methods in detecting hepatitis C virus (HCV) in the global population. Using various data parameters, the research used machine learning classification methods like k-nearest neighbors, naïve Bayes, neural network, and random forest. The neural network method showed the highest accuracy at 95.12%, surpassing KNN, naïve Bayes, and RF with accuracies of 89.43%, 90.24%, and 94.31%, respectively.

In [52] the objective of this study was compared the prognostic capabilities of four machine learning models for determining hepatitis B and C virus susceptibility. The study, conducted in Romania from January to November 2022, used data from surveys to analyze support vector machine (SVM), random forest (RF), naïve Bayes (NB), and K nearest neighbors (KNN) models. All models showed improved performance in forecasting HCV status, with KNN achieving the highest accuracy at 98.1%. The accuracy of predictions for HBV status varied from 78.2% to 97.6%.

In [53] the authors developed an algorithm using machine learning to identify cases of viral hepatitis B in patients from Nigeria. A patient dataset consisting of Hepatitis B Surface Antigen (HBsAg) findings as well as regular clinical chemistry and hematology blood tests was analyzed using a recursive partitioning method called "trees" and Support Vector Machines (SVMs). The dataset had a sample size of 916 patients. A model for the prediction diagnosis of HBV infection was created using these algorithms. In terms of accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating curve, our support vector machine (SVM)-based infection diagnostic model is competitive with immunoassay; it achieved a 91% sensitivity rate, 72.6% specificity rate, 88.2% accuracy rate, and F1-score of 0.89.

In [54] the purpose of this study was over the past 30 years, public health efforts to reduce viral hepatitis have grown, with the public health impact becoming more apparent in 2010. Major gaps in responsiveness and increasing mortality exist. The five viruses that cause most viral hepatitis are A, B, C, D, and E. HBV, HCV, and HDV cause chronic hepatitis, which may lead to liver scarring and malignancy. Both HBV and HCV cause 96% of hepatitis virus fatalities. An expert system using Naive Bayes algorithm is being developed to identify the start of the illness, with test findings yielding 90% accuracy.

In [55] the researchers used the Correlated Naïve Bayes Algorithm was applied to a dataset of hepatitis patients, predicting their healing rates based on various attributes like age, gender, and medical history. The algorithm outperformed its predecessors in terms of accuracy, reaching 86.04%, 0.77 for precision, 0.83 for recall, and 0.79 for f1-score. The study was chosen for testing due to its ease of finding necessary libraries in the Python programming language. The results highlight the algorithm's effectiveness in analyzing data from hepatitis patients.

In [56] the authors focused on improving the performance of a model by removing outliers, addressing class imbalance, and extracting strongly linked characteristics. The mean/mode imputation approach was used to eliminate missing data, while an oversampling strategy and z-score approach were employed to identify outliers. The integrated feature selection method was used to identify strongly linked features. Classic ML algorithms like K-Nearest Neighbors, Naive Bayes, and Random Forest were used to determine hepatitis. The model was evaluated using a 10-fold cross validation approach, achieving the highest classification accuracy of 97.44% from Random Forest.

In [57] the authors proposed a method for forecasting hepatitis disease using data mining techniques. By replacing absent values with mean values, the prediction models are improved. The dataset is processed using nine algorithms to compute prediction accuracy. The study uses random search hyperparameter optimization to optimize performance. The results show a significant enhancement in model accuracy of 83.87% compared to the Random Search algorithm for neural networks, which achieved a peak score of 91.98%. This suggests that optimized models can achieve greater accuracy after hyperparameter tuning.

In [58] the authors focused Hepatitis C, a major public health issue, is often caused by viral infections, particularly Hepatitis C. Researchers have used the K-Nearest Neighbor (KNN) approach to predict Hepatitis C risks. However, the algorithm's performance can be hampered by inaccurate selection of the K value. To address this, Particle Swarm Optimization (PSO) was used to optimize the KNN value. The results showed that the KNN algorithm achieved an accuracy of 97.24% when K was set to 5 and 3, and the addition of PSO improved the accuracy by 2.07% to reach 99.31%. This highlights the effectiveness of PSO in improving KNN performance and establishing the model as a potential approach for classifying Hepatitis C illness.

In [59] the Intelligent Hepatitis C Stage Diagnosis System (IHSDS) uses ANN and machine learning to estimate a human's Hepatitis C Stage. The dataset, derived from the UCI machine learning repository, includes 29 features. 70% is used for training and 30% for validation. The study selected 19 most revered characteristics. The IHSDS achieved 98.89% training accuracy and 94.44% validation accuracy, outperforming previous models.

In [60] the authors explored various data mining techniques for diagnosing hepatitis, comparing their accuracy and training time. The authors used Naive Bayes, K-nearest neighbor, and Random Forest classifiers in the WEKA program. Naïve Bayes is used for text categorization, while K-nearest neighbor is a simple, supervised machine learning technique applicable to regression and classification problems. Random Forest is widely used due to its simplicity and ability to be applied to both tasks. Naïve Bayes achieved 93.2% accuracy, while Random Forest achieved 98.6% with 10-fold cross-validation. K-nearest neighbor achieved 95.8% accuracy using the same cross-validation technique.

Table 1. Overview of the literature on Hepatitis Diagnosis based on Machine Learning Classification Algorithms

Ref	Year	Datasets	Algorithm	Advantages	Limitations	Result
[32]	2023	859 patients and includes 12 different features in Egypt	Naive Bayes, random forest (RF), K-nearest neighbor, and logistic regression	The proposed framework presented a robust, reliable, and efficient approach for predicting the Hepatitis C Virus among healthcare workers in Egypt, offering improved accuracy through feature selection and parameter tuning.	include a small dataset size, specific sample representation, limited features, and potential efficiency concerns in model building.	94.88%
[33]	2021	155 patients from the UCI	KNN, Naive Bayes, SVM, MLP and RF	This approach offers valuable insights for healthcare practitioners aiming to diagnose hepatitis effectively.	Larger dataset and diverse algorithms are needed for improved accuracy.	92.41%
[34]	2021	416 patients from Indian	KNN and Naïve Bayes	Overall, this approach offers valuable insights into predictive healthcare analytics, aiming to improve diagnosis and treatment outcomes for liver disease.	small dataset, few variables, and algorithm choices. Future research needs larger datasets and diverse algorithms for better accuracy.	72.5% and 63.19
[35]	2021	615 in the UCI library	decision tree Multi-layer perceptron, Bayesian network, and	Considering the superior performance of the provided methodology compared to all previously recommended ways in prior research, the proposed system in this study may be effectively used for diagnosing HCV and assessing its severity.	need more performance metrics, information about dataset, poor generalizability	99.9%
[36]	2021	113 data and 5 features	SVM and random forest	Prediction using SVM and Random Forest. SVM classifies using discriminant hyperplanes, whereas Random Forest uses random variables. The research compares two techniques to determine their hepatitis data prediction accuracy.	include a small dataset size of 113 data points, which may limit result generalizability. essential performance metrics like sensitivity and specificity were not included.	99.13% and 98.43%

[37]	2020	155 in the UCI library	Naïve Bayes and KNN	This approach underscores the importance of leveraging existing data and refining classification algorithms for more reliable predictive outcomes in healthcare and other domain	using 155 data points and focus on only 19 attributes. they tacked data only from the UCI that limit the study's scope, used only two algorithms maybe overlook other effective methods.	74.19% and 54.84%
[38]	2022	155 in the UCI	RF, DT, and SVM	showed a robust correlation between traits and the diagnosis of hepatitis, potentially improving early diagnosis and reducing acute effects.	Limited dataset size may affect model accuracy.	96.1%, 94.3%, and 92.2%
[39]	2023	155 in the UCI	SVM, LR, KNN and RF	Machine learning approaches improve hepatitis diagnosis by assessing classifier performance on UCI dataset.	Limited discussion on implementation challenges and features.	93.18%
[40]	2020	155 in the UCI	SVM, KNN, and ANN	Compared machine learning tools and neural networks for diagnosing and predicting life expectancy of Hepatitis patients, focusing on accuracy rate and mean square error.	Small dataset, biased data, limited algorithm focus.	96%
[41]	2022	44 samples	K-NN and DT	Suggested that employing K-NN classifiers may lead to optimized resource allocation in healthcare settings, potentially improving efficiency and reducing costs associated with hepatitis C detection.	small sample, bias, algorithm dependency, generalizability, resource constraints, publication bias.	0.42%
[42]	2021	73 in the UCI	KNN, naïve Bayes, NN, and RF	Compared the accuracy, sensitivity, and specificity of K-Nearest Neighbor Classifier and Decision Tree classifiers in detecting Novel Hepatitis C.	Limited sample size affects generalizability of results	89.43%, 90.24%, and 94.31%.
[43]	2023	1359 participants	SVM, RF, NB, and KNN.	Provided significant advantages in the field of hepatitis B (HBV)	small patient cohort, limited predictors	8.1%, 7.6%, 95.7%,

				and hepatitis C (HCV) infection prediction. By systematically evaluating and comparing the predictive performances of four machine learning.	affecting comprehensive predictive performance assessment.	and 97.6%
[44]	2023	916 participants	trees and SVM	Proposed approach has the potential to significantly improve clinical management and patient outcomes in resource-limited settings.	A small sample size. Exclusion criteria focused on HBV patients without co-infections and variables with high missing data. Machine learning model performance may vary across populations.	85.4%
[45]	2020	10 sample	Naive Bayes	Overall, the mobile-based diagnostic system holds significant potential in improving healthcare accessibility and outcomes for individuals at risk of hepatitis infection.	Testing accuracy based on only 10 sample data, focus on hepatitis A, B, and C may not cover all possible liver diseases. expert system dependency.	90%
[46]	2023	142 in the UCI	Naïve Bayes	The study highlights age and medical history as key factors influencing treatment choices and patient outcomes, but acknowledges limitations in attribute independence for improved healthcare accuracy.	small dataset (142 records), public data reliance, no algorithm comparison, and lacking data preprocessing details.	86.04%
[47]	2024	123 from the University of California	KNN, NB and RF	This approach enables timely identification of hepatitis cases, facilitating prompt treatment and improved patient outcomes.	focuses on early hepatitis detection using machine learning, with a limitation in sample size and demographic scope	97.44%
[48]	2023	155 in the UCI	K-NN, NB, SVM, ML and RF, Gradient	To improve classification model accuracy, we have shown that missing value datasets and	a small dataset impacting representation and limited hyperparameter	83.87%,

			Boosting, and K-Means	feature selection strategies are critical. This method may improve classification model accuracy and dependability, enabling better decision-making in numerous fields.	optimization, potentially affecting model performance and feature selection challenges due to complexities in highly correlated relationships among features.	
[49]	2023	589 samples	KNN	Using PSO on K-Nearest Neighbor (KNN) to diagnose Hepatitis C has several benefits. In extreme circumstances, hepatitis C may cause liver inflammation and death. Early forecasts can raise awareness and fight this menace.	Uncertain K value in KNN impacts accuracy. PSO optimizes K for improved classification. Further research may explore additional factors for enhanced performance.	97.24%
[50]	2021	969 samples	NN	Use of machine learning algorithms in illness prediction accelerates diagnosis and assures prompt treatment, increasing patient outcomes.	a small sample size of 22 participants and focused only on K-NN and decision tree classifiers, potentially overlooking other algorithms and factors that could influence the results.	94.44%
[51]	2019	Clinical in rural Bangladesh	NB, KNN and RF	Showed different data mining algorithms for diagnosing hepatitis and their accuracy and training duration.	the exclusion of hepatitis D due to testing constraints and the use of only a few classifier algorithms and interfaces in the Weka tool.	89.43%, 90.24%, and 94.31%

C. Result and Discussion

After doing a thorough analysis of several research on the prediction and diagnosis of Hepatitis using machine learning algorithms, it is evident that Random Forest constantly demonstrates superior performance and consistently ranks among the top-performing algorithms. Multiple studies, including [35], [38], [39], [44], [47], and [51], have consistently shown that Random Forest has good accuracy rates in the identification and classification of hepatitis patients. This technique has resilient performance across many datasets and experimental configurations, often surpassing other algorithms like as Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and K-nearest neighbor. Random Forest is very versatile and excellent in dealing with complex datasets and identifying significant patterns, which is why it is often the chosen method for diagnosing Hepatitis. In addition, several studies also emphasize the effectiveness of Support Vector Machines (SVM) and K-nearest neighbor (KNN) in obtaining high levels of accuracy, suggesting that they have promise for future investigation in the diagnosis of hepatitis.

D. Conclusion

Hepatitis stands out as a prevalent global affliction affecting people worldwide, underscoring the crucial significance of advancements in early detection and prognosis for liver diseases to ensure overall well-being. This paper focuses on exploring prevalent machine learning techniques aimed at classifying hepatitis, highlighting their relevance in enhancing diagnostic capabilities for this condition. Through a systematic examination of various classification algorithms such as Support Vector Machines, K-nearest Neighbors, Random Forest, and Naive Bayes, this review highlights the diverse array of tools available for healthcare practitioners in the diagnosis and management of hepatitis. A comparative examination was conducted using filter-based feature selection algorithms to categorize hepatitis disease. The evaluation of these feature selection methods was based on performance metrics, revealing that Random Forest emerges as the most precise, dependable, and easily detectable option.

E. References

[1] S. Patel, "Hepatitis," *SA Pharmaceutical Journal,* vol. 82, no. 6, pp. 20-23, 2015.

[2] H. Dwivedi and G. S. Rathore, *Virus Hepatitis (World Monster)*. Walnut Publication, 2020.

[3] A. John, Z. G. Ibrahim, S. Y. Magaji, J. S. Ede, and S. A. Bello, "Hepatitis B and C: a twin silent killer," *World J Pharm Res,* vol. 3, 2018.

[4] N. Ali, *Understanding Hepatitis: An Introduction for Patients and Caregivers*. Rowman & Littlefield, 2018.

[5] S. Khan *et al.*, "Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning," *Photodiagnosis and photodynamic therapy,* vol. 23, pp. 89-93, 2018.

[6] B. T. Chicho, A. M. Abdulazeez, D. Q. Zeebaree, and D. A. Zebari, "Machine learning classifiers based classification for IRIS recognition," *Qubahan Academic Journal,* vol. 1, no. 2, pp. 106-118, 2021.

[7] S. C. Nandipati, C. XinYing, and K. K. Wah, "Hepatitis C virus (HCV) prediction by machine learning techniques," *Applications of modelling and simulation,* vol. 4, pp. 89-100, 2020.

[8] K. Moulaei, H. Sharifi, K. Bahaadinbeigy, A. A. Haghdoostd, and N. Nasiri, "Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis," *International Journal of Medical Informatics,* p. 105243, 2023.

[9] S. M. Abas and A. M. Abdulazeez, "Detection and Classification of Leukocytes in Leukemia using YOLOv2 with CNN," *Asian Journal of Research in Computer Science,* vol. 8, no. 3, pp. 64-75, 2021.

[10] B. Karlik, "Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers," *Journal of Science and Technology,* vol. 1, no. 1, pp. 49-62, 2011.

[11] H. Wang, Y. Liu, and W. Huang, "Random forest and Bayesian prediction for Hepatitis B virus reactivation," in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2017: IEEE, pp. 2060-2064.

[12] M. A. Remita, A. Halioui, A. A. Malick Diouara, B. Daigle, G. Kiani, and A. B. Diallo, "A machine learning approach for viral genome classification," *BMC bioinformatics,* vol. 18, pp. 1-11, 2017.

[13] A. E. Mehyadin and A. M. Abdulazeez, "Classification based on semi-supervised learning: A review," *Iraqi Journal for Computers and Informatics,* vol. 47, no. 1, pp. 1-11, 2021.

[14] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b,* vol. 4, no. 51-62, p. 56, 2017.

[15] Z.-H. Zhou, *Machine learning*. Springer nature, 2021.

[16] A. J. Mueller-Breckenridge *et al.*, "Machine-learning based patient classification using Hepatitis B virus full-length genome quasispecies from Asian and European cohorts," *Scientific Reports,* vol. 9, no. 1, p. 18892, 2019.

[17] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and region growing for breast cancer segmentation," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 2019: IEEE, pp. 88-93.

[18] D. Castaneda, A. J. Gonzalez, M. Alomari, K. Tandon, and X. B. Zervos, "From hepatitis A to E: A critical review of viral hepatitis," *World journal of gastroenterology,* vol. 27, no. 16, p. 1691, 2021.

[19] F. Maulidina, A. R. Laeli, and Z. Rustam, "Comparison Between Logistic Regression and Support Vector Machine for Hepatitis Classification," in *3RD INTERNATIONAL CONFERENCE ON MATHEMATICAL AND RELATED SCIENCES: CURRENT TRENDS AND DEVELOPMENTS PROCEEDINGS BOOK*, 2020, p. 55.

[20] D. M. Ahmed, A. M. Abdulazeez, D. Q. Zeebaree, and F. Y. Ahmed, "Predicting university's students performance based on machine learning techniques," in *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*, 2021: IEEE, pp. 276-281.

[21] A. M. Richardson and B. A. Lidbury, "Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data," *BMC bioinformatics,* vol. 14, pp. 1-8, 2013.

[22] D. Q. Zeebaree, D. A. Hasan, A. M. Abdulazeez, F. Y. Ahmed, and R. T. Hasan, "Machine Learning Semi-Supervised Algorithms for Gene Selection: A

Review," in *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, 2021: IEEE, pp. 165-170.

[23] M. D. Genemo, "Diagnosis of Hepatitis using Supervised Learning algorithm," *Indonesian Journal of Data and Science (IJODAS) ISSN,* vol. 2715, p. 9930, 2023.

[24] S. Hashem *et al.*, "Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease," *Computer methods and programs in biomedicine,* vol. 196, p. 105551, 2020.

[25] J.-Y. Zhou, L.-W. Song, R. Yuan, X.-P. Lu, and G.-Q. Wang, "Prediction of hepatic inflammation in chronic hepatitis B patients with a random forest-backward feature elimination algorithm," *World journal of gastroenterology,* vol. 27, no. 21, p. 2910, 2021.

[26] M. H. Alsharif, A. H. Kelechi, K. Yahya, and S. A. Chaudhry, "Machine learning algorithms for smart data analysis in internet of things environment: taxonomies and research trends," *Symmetry,* vol. 12, no. 1, p. 88, 2020.

[27] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends,* vol. 2, no. 01, pp. 20-28, 2021.

[28] H.-L. Chen, D.-Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," *Expert systems with applications,* vol. 38, no. 9, pp. 11796-11803, 2011.

[29] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Computer methods and programs in biomedicine,* vol. 108, no. 2, pp. 570-579, 2012.

[30] Y. Chen *et al.*, "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B," *Computers in biology and medicine,* vol. 89, pp. 18-23, 2017.

[31] A. Singh and B. Pandey, "Diagnosis of liver disease using correlation distance metric based k-nearest neighbor approach," in *Intelligent Systems Technologies and Applications 2016*, 2016: Springer, pp. 845-856.

[32] S. L. Sudhakaran, D. Madathil, M. Arumugam, and V. Sundararajan, "Drug Development for Hepatitis C Virus Infection: Machine Learning Applications," *Global Virology III: Virology in the 21st Century,* pp. 117-129, 2019.

[33] P. Kumar and R. S. Thakur, "Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach," *Multimedia Tools and Applications,* vol. 80, pp. 16515-16535, 2021.

[34] R. Shiomi, H. Shimasaki, H. Takano, and H. Taoka, "A study on operating lifetime estimation for electrical components in power grids on the basis of analysis of maintenance records," *Journal of International Council on Electrical Engineering,* vol. 9, no. 1, pp. 45-52, 2019.

[35] 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.

[36] E. Dritsas and M. Trigka, "Supervised machine learning models for liver disease risk prediction," *Computers,* vol. 12, no. 1, p. 19, 2023.

[37] T. Karthikeyan and P. Thangaraju, "Best first and greedy search based CFS-Naïve Bayes classification algorithms for hepatitis diagnosis," *Biosciences and Biotechnology Research Asia,* vol. 12, no. 1, pp. 983-990, 2015.

[38] M. Ashraf, G. Chetty, D. Tran, and D. Sharma, "Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based feature selection and Naïve bayes," in *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part IV 19*, 2012: Springer, pp. 272-280.

[39] S. Vijayarani and S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms," *International Journal of Science, Engineering and Technology Research (IJSETR),* vol. 4, no. 4, pp. 816-820, 2015.

[40] A. Choi, N. Tavabi, and A. Darwiche, "Structured features in naive Bayes classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, vol. 30, no. 1.

[41] 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," *Knowledge and Information Systems,* vol. 65, no. 6, pp. 2595-2617, 2023.

[42] M. J. Nayeem, S. Rana, F. Alam, and M. A. Rahman, "Prediction of hepatitis disease using K-nearest neighbors, Naive Bayes, support vector machine, multi-layer perceptron and random forest," in *2021 international conference on information and communication technology for sustainable development (ICICT4SD)*, 2021: IEEE, pp. 280-284.

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

[44] A. Orooji and F. Kermani, "Machine learning based methods for handling imbalanced data in hepatitis diagnosis," *Frontiers in Health Informatics,* vol. 10, no. 1, p. 57, 2021.

[45] J. E. Aurelia, Z. Rustam, I. Wirasati, S. Hartini, and G. S. Saragih, "Hepatitis classification using support vector machines and random forest," *IAES International Journal of Artificial Intelligence,* vol. 10, no. 2, p. 446, 2021.

[46] R. Alfyani, "Comparison of Naïve Bayes and KNN algorithms to understand hepatitis," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2020: IEEE, pp. 196-201.

[47] I. I. Ahmed, D. Y. Mohammed, and K. A. Zidan, "Diagnosis of hepatitis disease using machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 26, no. 3, pp. 1564-1572, 2022.

[48] R. K. Sachdeva, P. Bathla, P. Rani, V. Solanki, and R. Ahuja, "A systematic method for diagnosis of hepatitis disease using machine learning," *Innovations in Systems and Software Engineering,* vol. 19, no. 1, pp. 71-80, 2023.

[49] V. K. Yarasuri, G. K. Indukuri, and A. K. Nair, "Prediction of hepatitis disease using machine learning technique," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2019: IEEE, pp. 265-269.

[50] D. Sravanthi and D. J. Rani, "Comparative Analysis of Hepatitis C Using K-Nearest Neighbor Classifier and Decision Tree Classifier," *Cardiometry,* no. 25, pp. 1010-1016, 2022.

[51] L. Syafaâ, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of machine learning classification methods in hepatitis C virus," *Jurnal Online Informatika,* vol. 6, no. 1, pp. 73-78, 2021.

[52] V. Harabor *et al.*, "Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity," *International journal of environmental research and public health,* vol. 20, no. 3, p. 2380, 2023.

[53] B. I. Ajuwon *et al.*, "The development of a machine learning algorithm for early detection of viral hepatitis B infection in Nigerian patients," *Scientific Reports,* vol. 13, no. 1, p. 3244, 2023.

[54] G. Irfansyah, U. Darusallam, and B. Benrahman, "Early Diagnosis Expert System Hepatitis Using Naive Bayes Method: Early Diagnosis Expert System Hepatitis Using Naive Bayes Method," *Jurnal Mantik,* vol. 3, no. 4, pp. 182-187, 2020.

[55] Y. Yulhendri, M. Malabay, and K. Kartini, "Correlated Naïve Bayes Algorithm to Determine Healing Rate of Hepatitis Patients," *International Journal of Science, Technology & Management,* vol. 4, no. 2, pp. 401-410, 2023.

[56] M. Mim, S. Akter, M. J. Nayeem, S. Rana, and M. R. Islam, "A Predictive Approach for Hepatitis Disease Diagnosis in Early Stage Using Machine Learning Techniques," *Julker and Rana, Sohel and Islam, Md. Rabiul, A Predictive Approach for Hepatitis Disease Diagnosis in Early Stage Using Machine Learning Techniques*.

[57] H. D. Saputra, A. I. E. Efendi, E. Rudini, D. Riana, and A. S. Hewiz, "Hepatitis Prediction Using K-NN, Naive Bayes, Support Vector Machine, Multilayer Perceptron and Random Forest, Gradient Boosting, K-Means," *Journal Medical Informatics Technology,* pp. 96-100, 2023.

[58] S. Setianingsih, M. U. Chasanah, Y. I. Kurniawan, and L. Afuan, "IMPLEMENTATION OF PARTICLE SWARM OPTIMIZATION IN K-NEAREST NEIGHBOR ALGORITHM AS OPTIMIZATION HEPATITIS C CLASSIFICATION," *Jurnal Teknik Informatika (Jutif),* vol. 4, no. 2, pp. 457-465, 2023.

[59] M. B. Butt *et al.*, "Diagnosing the stage of hepatitis C using machine learning," *Journal of Healthcare Engineering,* vol. 2021, p. 8062410, 2021.

[60] 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)*, 2019: IEEE, pp. 1-7.