International Journal of Contemporary Mathematical Sciences
Vol. 21, 2026, no. 1, 11 - 29
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ijcms.2026.92018

Diagnosing Hepatitis B Disease Using New Dimensional Space Partitioning Model Based on Hyperplanes

J.C. Martínez-Perales

Instituto Politécnico Nacional. Escuela Superior de Cómputo.
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico

B. Luna-Benoso

Instituto Politécnico Nacional. Escuela Superior de Cómputo
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico

J. Cortés-Galicia

Instituto Politécnico Nacional. Escuela Superior de Cómputo.
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico

U. S. Morales-Rodríguez

Instituto Politécnico Nacional. Escuela Superior de Cómputo.
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico

Abstract

Hepatitis B is a liver infection caused by the hepatitis B virus and is considered a global health issue. Although traditional hepatitis B detection is performed through blood tests, these conventional methods in healthcare are increasingly being replaced or supplemented by Artificial Intelligence (AI) models. Machine Learning, a field of AI, enables computers to learn from data sets. Machine Learning has been applied in the healthcare sector, assisting medical experts in decision-making and yielding favorable results. This study proposes diagnosing hepatitis B through a new classification model that uses hyperplanes to partition the dimensional space into sections where vectors belong to the same class. This partitioning infers a decision tree, allowing for the diagnosis of hepatitis B. The proposed model was applied to the hepatitis dataset from the University of California, Irvine (UCI) Machine Learning Repository and was compared with other state-of-the-art classification models. The performance evaluation metrics used were Sensitivity, Specificity, and Accuracy. The evaluation metrics obtained with the proposed model were 71%, 96%, and 90.63% for Sensitivity, Specificity, and Accuracy, respectively, achieving superior results compared to the models with which it was compared.

Keywords: Decision tree, Gini index, hepatitis B, hyperplanes, machine learning

I. Introduction

Hepatitis B is an infectious disease caused by the hepatitis B virus (HBV), with high mortality and morbidity rates. In 2015 alone, around 257 million people were chronically infected with HBV [1], and in 2019, approximately 296 million people worldwide [2]. It is estimated that around 1.5 million people are infected each year [3], leading to more than 1.1 million deaths annually [4]. Symptoms in people infected with hepatitis B can include stomach pain, fever, nausea, vomiting, jaundice, yellowing of the eyes, dark urine, abdominal pain, weakness, fatigue, and joint pain [5]. Hepatitis B can lead to severe conditions, such as hepatocellular carcinoma, liver cirrhosis, liver cancer, liver failure, or kidney diseases, among others [6-9]. Hepatitis B transmission occurs through blood transfusion or contact, mother-to-child transmission during birth, liver transplants, syringe and needle reuse, sexual transmission, and through contact with bodily fluids like saliva and secretory skin lesions [10-11]. In 2011, the World Health Organization (WHO) recognized viral hepatitis as a global threat, and in 2016, the World Health Assembly approved a Global Health Sector Strategy (GHSS) aiming to reduce morbidity and mortality rates caused by viral hepatitis by 2030. The strategy targets a 90% reduction in new infections, a 65% reduction in mortality rate, a 90% increase in diagnostics, and an 80% increase in treatment rates [12].

Artificial Intelligence (AI) in healthcare has emerged as a promising tool for early disease detection, supporting physicians in clinical decision-making [13].

Such is the case for AI applied to hepatitis B detection. Although hepatitis B diagnosis is typically performed through blood analysis, many affected populations lack access to clinical blood tests [14]. AI offers an alternative to traditional hepatitis B detection by processing large datasets that provide a comprehensive view of regional epidemiological profiles [15]. AI leverages Machine Learning (ML) models to determine if a patient has hepatitis B based on patient records showing symptoms of the disease.

ML methods and models applied to early hepatitis detection help minimize diagnostic errors and reduce processing time [16]. Other works such as in [17] use various ML models such as eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) to identify the model yielding the best results based on specific demo-graphic and laboratory information. While many studies use classic ML models like K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) for hepatitis B predictions [18-19], some improve classic models to make them more robust or propose new ML models. For example in [20], applies a Penalized Dis-similarity Measure (PDM) approach, integrating a PDM called Feature Weighted Penalty-based Dis-similarity (FWPD) into K-NN, forming the K-NN-FWPD classifier, which can directly handle datasets with missing features without preprocessing. In another study [21], proposes a hybrid model between SVM and Simulated Annealing (SA), while that in [22] proposes an intelligent system to diagnose hepatitis using a variant of SVM: the Least Square Support Vector Machine Classifier (LSSVM). Developing new ML models is essential due to the No-Free-Lunch Theorem (NFLT), which implies that no universally optimal model exists; in other words, no classifier achieves the best results for all problems. Thus, creating new machine learning models to achieve competitive results on specific problems is crucial. Studies such as [15] utilize prospective cohort screening studies in adults undergoing viral hepatitis testing in family clinics and evaluate ML algorithms like SVM, RF, Naïve Bayes Classifier (NB), and K-NN on this data. Other studies such as in [23] use patient medical records, laboratory clinical data, demographic characteristics, routine blood tests, liver biochemistry, coagulation, and virology data to form training and test datasets, applying a series of ML algorithms. Several studies have used ML models to predict hepatitis B on clinical datasets, often utilizing the hepatitis dataset from the University of California, Irvine (UCI) Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/hepatitis) [21]. In pattern recognition, in addition to feature extraction and classification, feature selection can sometimes be advantageous for choosing relevant variables. For example, in [24], the authors compare different classifiers with various filter-based feature selection methods. The models used include NB, LR, and J48 Decision Tree, with feature selection methods such as Cfs Subset Eval, Info Gain Attribute Eval, and Principal Component Analysis (PCA). The authors concluded that without applying any feature selection filters, NB achieved the best results with an accuracy of 84.51%. However, when using PCA, NB obtained the best result with an accuracy of 88.38%. In [25], the authors propose three feature selection methods: PCA, Factor Analysis (FA), and Linear Discriminant Analysis (LDA), applied to an ensemble

voting classifier using LR, RF, K-NN, SVM, and MultiLayer Perceptron (MLP) classifiers. This study achieved an accuracy of 88.10%. Early hepatitis B detection methods and models are not limited to conventional ML models, as some studies use Deep Learning (DL) through deep layers in a Deep Neural Network (DNN) [26]. While these models achieve good results, they are often black-box models, where the results are challenging to interpret and rely on complex mathematical systems.

This study proposes a new machine learning model that, by partitioning dimensional space with hyperplanes, constructs a decision tree for detecting hepatitis B in patients. Generally, decision tree models are easy to interpret, as the flow resembles a flowchart where outcomes depend on a series of decisions made at each node. Additionally, the model's construction allows a clear visualization of how the dimensional space is divided into classes. The UCI Machine Learning Repository dataset was used to apply the pro-posed model, and the results achieved competitive accuracy compared to other ML studies with which it was benchmarked.

II. Methodology

II.1 Gini Index

The Gini index G(e) of a node e is defined as $G(e) = 1 - \sum_I P(i|e)^2$ where $P(i|e)^2$ is conditional probability of category i in node e. The Gini index is a measure that calculates the degree of impurity of a node in a decision tree, indicating the homogeneity of the dimensional space partition inferred by the decision tree. When the Gini index is equal to zero, it indicates that the nodes are pure; in other words, all vectors in the set belong to the same class. On the other hand, values greater than zero and up to one indicate nodes with impurities, meaning that the vectors in the set belong to different classes [27].

II.2 Missing Data

Data incompleteness is a major challenge when using ML, and although several works prefer to discard incomplete records [28], it is not always the most appropriate, since valuable information can be lost. For this, there are alternatives such as K-NN imputation, which consists of estimating missing data using K-NN, that is, it predicts missing data based on observations of the values of the closest records in the feature space [29].

II.3 Feature Selection

Feature selection is a technique used in ML that consists of determining the set of features or attributes that are most relevant to the performance of a model, eliminating those features that contribute little or nothing. There are several methods among PCA. PCA is a technique that allows dimensionality reduction by

using data that have correlated features in a new set of uncorrelated linear variables called principal components [30].

II.4 Proposed Model

The model presented here takes as reference the training set $TS = \{(\bar{x}_1, C_1), \dots, (\bar{x}_m, C_m)\}$ where $\bar{x}_i \in \mathbb{R}^n$ represents a feature vector, and C_i denotes its class, with $i = 1, \dots, m$. Although the proposed model is initially shown for a binary classification problem, it will later be explained how it can be extended to an n-class problem. Suppose each class in the training set corresponds to either class A or B. The equation of a hyperplane is given by $\bar{w}^t \bar{x} + b = 0$ where $\bar{w} = (w_1, w_2, \cdots, w_n)$ is the normal vector to the hyperplane, $\bar{x} = (x_1, x_2, \cdots, x_n)$ are the coordinates of the space and b the independent term or bias. The proposed model makes use of a series of hyperplanes that are built from the average of sets of vectors that belong to different classes, for its explanation, consider $\bar{P}_{\mathbb{A}}$ and $\bar{P}_{\mathbb{B}}$ represent the mean of the set of vectors belonging to class A and class B, respectively, defined as, $\bar{P}_{\mathbb{A}} = \frac{\sum_{i \in \mathbb{A}} \bar{x}_i}{n_{\mathbb{A}}}$ and $\bar{P}_{\mathbb{B}} = \frac{\sum_{i \in \mathbb{B}} \bar{x}_i}{n_{\mathbb{B}}}$ where $n_{\mathbb{A}}$ and $n_{\mathbb{B}}$ represent the number of vectors belonging to class A and class B, respectively. In cases where $\bar{P}_{\mathbb{A}} = \bar{P}_{\mathbb{B}}$, one vector from class A is removed, and $\bar{P}_{\mathbb{A}}$. is recalculated. Let $r = \bar{P}_{\mathbb{A}} + \alpha \bar{P}_{\mathbb{B}}$ denotes the parametric equation of the line passing through $\bar{P}_{\mathbb{A}}$ and $\bar{P}_{\mathbb{B}}$. Fig. 1 shows a set of points in space, where the red points belong to class A, and the blue points belong to class B, with the line connecting points $\bar{P}_{\mathbb{A}}$ and $\bar{P}_{\mathbb{B}}$ also depicted.

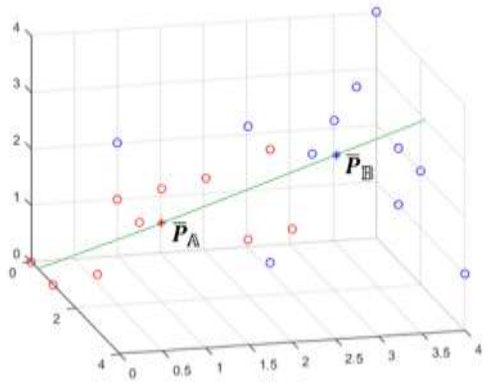


Figure 1. Line joining points $\bar{P}_{\mathbb{A}}$ and $\bar{P}_{\mathbb{B}}$ corresponding to the mean of the red and blue points, respectively

Once the parametric equation of the line joining $\bar{P}_{\mathbb{A}}$ and $\bar{P}_{\mathbb{B}}$ is established, for each point $\bar{x}_i$ in the training set, the equation of the hyperplane is calculated ${\bar{w}_i}^t \bar{x} + b_i = 0$ that passes by $\bar{x}_i$ and is perpendicular to the vector $\overline{P_{\mathbb{A}} P_{\mathbb{B}}} = \bar{P}_{\mathbb{A}} - \bar{P}_{\mathbb{B}}$. Fig. 2 illustrates the hyperplane that passes through the training set point $\bar{x}_i$ and is perpendicular to the vector $\overline{P_{\mathbb{A}} P_{\mathbb{B}}}$.

Define $\bar{P}_{min_{\mathbb{A}}} = \min\{d(\bar{P}_{\mathbb{A}}, \overline{Pr}_{\bar{x}_i}): \bar{x}_i \in \mathbb{B}\}$ and $\bar{P}_{min_{\mathbb{B}}} = \min\{d(\bar{P}_{\mathbb{B}}, \overline{Pr}_{\bar{x}_i}): \bar{x}_i \in \mathbb{A}\}$. Here, $\bar{P}_{min_{\mathbb{A}}}$ is the orthogonal projection corresponding to the vectors associated with class B closest to $\bar{P}_{\mathbb{A}}$, and $\bar{P}_{min_{\mathbb{B}}}$ is the orthogonal projection corresponding to the vectors associated with class A closest to $\bar{P}_{\mathbb{B}}$. Once vectors $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$ are calculated, we determine the hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0, what is perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$ and passing through the midpoint between $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$. Fig.3 illustrates vectors $\bar{P}_{\mathbb{A}}$, $\bar{P}_{\mathbb{B}}$, $\bar{P}_{min_{\mathbb{A}}}$, $\bar{P}_{min_{\mathbb{B}}}$, and the hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0, which passes through the midpoint between $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$ and is perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$.

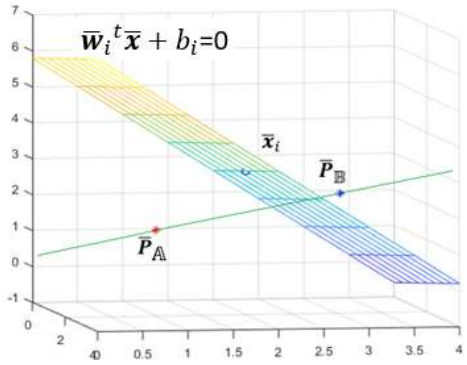


Figure 2. Hyperplane $\bar{w}_i{}^t\bar{x} + b_i$=0 passing through vector $\bar{x}_i$ and perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$

Once the hyperplane $\bar{w}_i{}^t\bar{x} + b_i = 0$ is obtained, the orthogonal projection of the vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$ onto $\bar{w}_i{}^t\bar{x} + b_i$=0 is calculated. Let $\overline{Pr_i}$ be the orthogonal projection of the vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$ onto the hyperplane $\bar{w}_i{}^t\bar{x} + b_i$=0. For each vector $\bar{x}_i$ in the training set, there exists an association with the respective orthogonal projection $\overline{Pr_i}$ to the hyperplane $\bar{w}_i{}^t\bar{x} + b_i$=0. These orthogonal projections lie on the line connecting $\bar{P}_{\mathbb{A}}$with $\bar{P}_{\mathbb{B}}$.

Define $\bar{P}_{min_{\mathbb{A}}} = \min\{d(\bar{P}_{\mathbb{A}}, \overline{Pr}_{\bar{x}_i}): \bar{x}_i \in \mathbb{B}\}$ and $\bar{P}_{min_{\mathbb{B}}} = \min\{d(\bar{P}_{\mathbb{B}}, \overline{Pr}_{\bar{x}_i}): \bar{x}_i \in \mathbb{A}\}$. Here, $\bar{P}_{min_{\mathbb{A}}}$ is the orthogonal projection corresponding to the vectors associated with class B closest to $\bar{P}_{\mathbb{A}}$, and $\bar{P}_{min_{\mathbb{B}}}$ is the orthogonal projection corresponding to the vectors associated with class A closest to $\bar{P}_{\mathbb{B}}$. Once vectors $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$ are calculated, we determine the hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0, what is perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$ and passing through the midpoint between $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$. Fig.3 illustrates vectors $\bar{P}_{\mathbb{A}}$, $\bar{P}_{\mathbb{B}}$, $\bar{P}_{min_{\mathbb{A}}}$, $\bar{P}_{min_{\mathbb{B}}}$,

and the hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0, which passes through the midpoint between $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$ and is perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$.

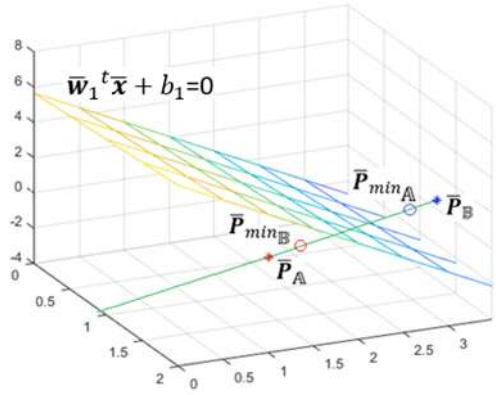
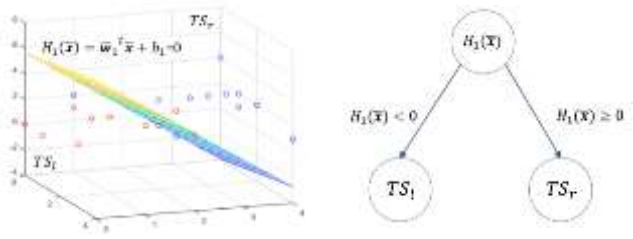


Figure 3. Hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0 passing through the midpoint between $\bar{P}_{min_{\mathbb{A}}}$ and $\bar{P}_{min_{\mathbb{B}}}$ and perpendicular to vector $\overline{P_{\mathbb{A}}P_{\mathbb{B}}}$

The hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0 divides the training set TS into two subsets corresponding to two sections in the dimensional space: $TS_l = \{(\bar{x}_i, C_i) \in TS\colon \bar{w}_1{}^t\bar{x}_i + b_1 < 0\}$ and $TS_r = \{(\bar{x}_i, C_i) \in TS\colon \bar{w}_1{}^t\bar{x}_i + b_1 \geq 0\}$. The hyperplane $H_1(\bar{x}) = \bar{w}_1{}^t\bar{x} + b_1$=0 serves as the root of a decision tree, where the root node is labeled by $H_1(\bar{x})$. For each vector $\bar{x}_r$ of dimensional space, the decision-making process depends on whether $\bar{w}_1{}^t\bar{x}_r + b_1 < 0$ or $\bar{w}_1{}^t\bar{x}_r + b_1 \geq 0$. Fig. 4 illustrates the dimensional space divided by the hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0 and the corresponding decision tree.



(**a**) (**b**)

Figure 4. (a) The hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0 divides the space into sections; (b) TS_1 and TS_2, thus inferring a decision tree

For each node in the tree labeled TS_1 and TS_2, corresponding to the sections given by hyperplane $\bar{w}_1{}^t\bar{x} + b_1$=0, the Gini index is calculated. If the Gini index in section TS_i $(i = l, r)$ is zero, the procedure stops in that section. Otherwise, if the Gini index in section TS_i $(i = l, r)$ has a non-zero value, the set TS is updated to the subset TS_i, and the procedure repeats until the Gini index in each subsequent subsection sequence equals zero. This implies that the subsections are fully homogeneous, meaning that all vectors in each subsection belong to the same class. The sequence of hyperplanes dividing the subsequent sections infers a decision tree, as shown in Fig. 5, concluding the training phase. For the testing phase, given an input vector $\bar{x}$ whose class needs to be determined, the classification depends on the value taken at each node when evaluating the corresponding hyperplane. Specifically, if $H_{TS_i}(\bar{x}) < 0$, the traversal goes to the left child; otherwise, it goes to the right child, where $H_{TS_i}(\bar{x}) = \bar{w}_{TS_i}{}^t\bar{x} + b_{TS_i} = 0$. Upon reaching a leaf node, it will be labeled with the class corresponding to the vector $\bar{x}$.

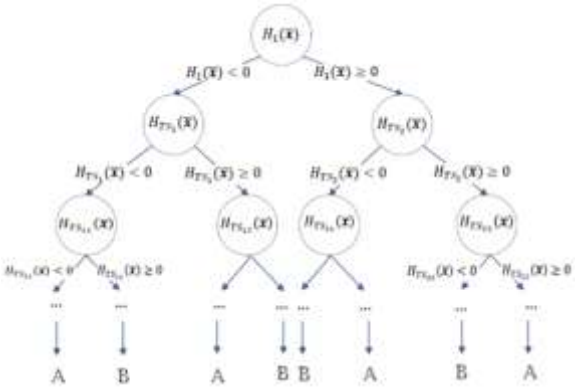


Figure 5: Decision tree obtained in the training phase, generated from the divisions of the hyperplanes resulting from subsequent subsection splits

Fig.6 illustrates the flowchart of the proposed model.

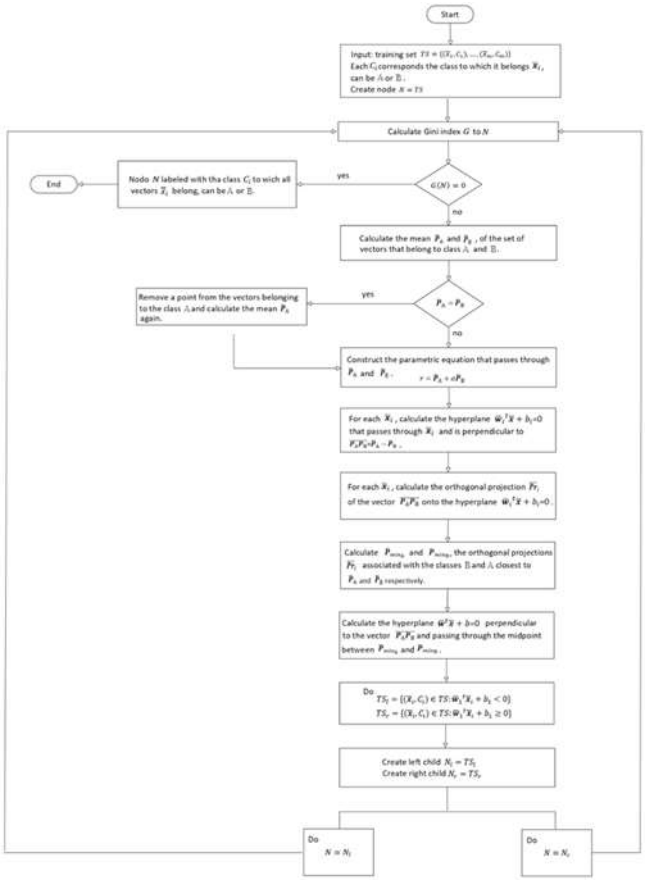


Figure 6. Flowchart of the proposed model

III. Experiments and Results

In this section, the experiments and results obtained by the proposed model are presented. First, the XOR logic function problem is addressed to observe the model's performance. Then, the hepatitis dataset from the UCI Machine Learning Repository is applied.

The XOR logic function problem was an issue that, in the early stages of Artificial Neural Networks (ANNs), the Simple Perceptron could not solve, and in general, it is a problem that cannot be solved using a linear classifier. Fig. 7 shows the table corresponding to the XOR function and the points plotted in space, where points (0,0) and (1,1) correspond to class $\mathbb{A} = 0$, and points (0,1) and (1,0) correspond to class $\mathbb{B} = 1$.

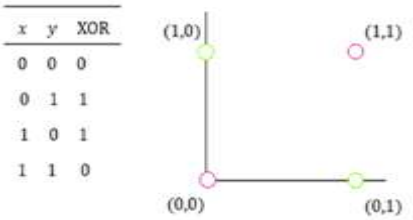
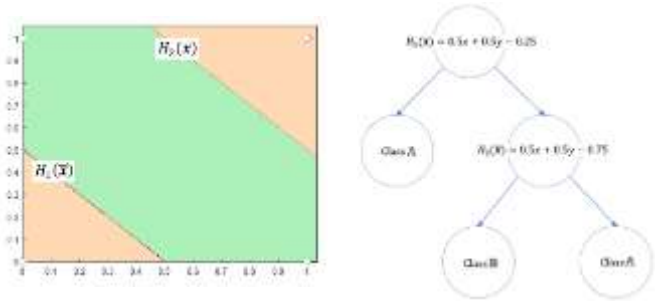


Figure 7. Function XOR

To apply the proposed model to the XOR logic function problem, we first define $TS = \{((0,0), \mathbb{A}), ((1,1), \mathbb{A}), ((0,1), \mathbb{B}), ((1,0), \mathbb{B})\}$. Next, the root node $N = TS$ is defined, and the Gini index is calculated, which in this case is $G(N) = 0.5$. Applying the flow diagram shown in Fig. 6 in a first cycle, the dimensional space is divided by the line equation $H_1(\bar{x}) = 0.5x + 0.5y - 0.25$. $H_1(\bar{x})$ divides the plane into two sections, with $TS_l = \{(\bar{x}_i, C_i) \in TS: H_1(\bar{x}_i) < 0\} = \{((0,0), \mathbb{A})\}$ and $TS_r = \{(\bar{x}_i, C_i) \in TS: H_1(\bar{x}_i) \geq 0\} = \{((1,1), \mathbb{A}), ((0,1), \mathbb{B}), ((1,0), \mathbb{B})\}$. We define $N_l = TS_l$ and $N_r = TS_r$, the left and right child nodes of node N, respectively. Since $G(N_l) = 0$, then node N_l is a leaf node with label class $\mathbb{A}$. In contrast, $G(N_r) = 0.44$, so by reapplying the flow diagram from Fig. 6 with $N = N_r$, section TS_r, divided by the line equation $H_2(\bar{x}) = 0.5x + 0.5y - 0.75$. Fig. 8 shows the dimensional plane division, and the decision tree generated by the proposed model.



(**a**) (**b**)

Figure 8. (a) Partition of the dimensional space using lines $H_1(\bar{x})$ and $H_2(\bar{x})$; (b)and the inferred decision tree

For the hepatitis B classification problem, the data was obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/hepatitis). The Josef Stefan Institute in former Yugoslavia donated this dataset. The dataset contains a total of 155 patient records divided into two classes: 32 "die" cases and 123 "live" cases. The 19 attributes are divided into 13 binary attributes and 6

attributes that contain between 6 and 8 distinct values. Approximately 48.30% of the data set includes missing values. Table 1 shows the 19 attributes that make up the data set and the percentage of missing values per attribute.

Table 1. The attributes of the hepatitis B disease dataset

The number of attributes	The name of attribute	The values of attribute	Missing percent %
1	Age	10, 20, 30, 40, 50, 60, 70, 80	0.00
2	Sex	Male, female	0.00
3	Steroid	Yes, No	1.00
4	Antivirals	Yes, No	0.00
5	Fatigue	Yes, No	1.00
6	Malaise	Yes, No	1.00
7	Anorexia	Yes, No	1.00
8	Liver big	Yes, No	6.00
9	Liver firm	Yes, No	7.00
10	Spleen palpable	Yes, No	3.00
11	Spiders	Yes, No	3.00
12	Ascites	Yes, No	3.00
13	Varices	Yes, No	3.00
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	4.00
15	Alk phosphate	33, 80, 120, 160, 200, 250	19.00
16	SGOT	13, 100, 200, 300, 400, 500	3.00
17	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	10.00
18	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90	43.00
19	HISTOLOGY	Yes, No	0.00
20	Class	Die, Alive	0.00

To manage the missing data in the dataset, experiments were conducted with data discarding and with K-NN imputation. Subsequently, attributes were selected using PCA. The best results were obtained by using K-NN imputation to handle the missing data and PCA for feature selection, obtaining a total of 10 selected attributes (1, 2, 5, 10, 11, 12, 13, 14, 17, 18). In its training phase, the proposed model generated a decision tree with 67 nodes, divided into 33 internal nodes and 34 leaf nodes, with a depth of 7. To evaluate the model's performance, a train-test split with 80% allocated to the training set and 20% was used, and the confusion matrix was generated. The confusion matrix describes how the real values are distributed concerning the values obtained by the model. Fig. 9 shows a confusion matrix where TP represents the true positives, TN the true negatives, FP the false positives, and FN the false negatives obtained with the classifier model.

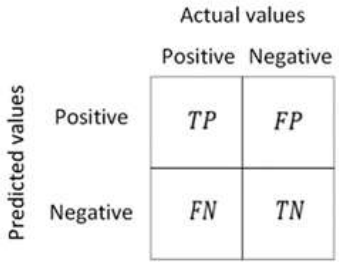


Figure 9. Confusion matrix

The performance evaluation metrics used in this study include Sensitivity, Specificity, and Accuracy. Sensitivity (SE) indicates the proportion of positive cases that the model correctly classifies, Specificity (SP) indicates the model's ability to predict negative cases, and Accuracy (ACC) corresponds to the proportion of correct predictions; that is, it measures how well a classification predicts a condition. See Eqs. (1)-(3) for the mathematical representation of the performance evaluation metrics:

$$SE = \frac{TP}{TP+FN} \times 100\%, \tag{1}$$

$$SE = \frac{TN}{TN+FP} \times 100\%, \tag{2}$$

$$ACC = \frac{TN+TP}{FN+FP+TN+TP} \times 100\%, \tag{3}$$

Fig.10 shows the confusion matrix obtained with the model using an 80%-20% train-test split of the dataset, where 80% was used for training and 20% for testing.

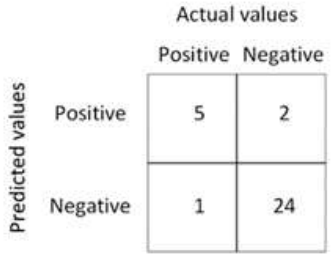


Figure 10. Results of the confusion matrix obtained by the proposed model

Table 2 shows a comparison of the proposed model with other models using the performance evaluation metrics (SE, SP, and ACC)

Table 2. Classification accuracy of the proposed model and other methods on hepatitis B disease

Model	ACC	SE	SP
SVM [31]	83.75	46.15	91.04
LR [31]	85.00	38.46	94.02
KNN [31]	81.25	46.15	88.05
RF [31]	85.00	38.46	94.02
DT [32]	73.00	66	80
LR [32]	82.00	74	89
SVM [32]	73.00	80	67
RF [32]	86.00	82	90
XGBoost [32]	90.00	88	92
SVM, Gaussian NB, LR, DT, KNN, and MLP [33]	87.00	---	---
KNN, LR, NB, DT, SVM, RF, Chi-square and Boruta [34]	90.32	---	---
CSFNN [35]	90.00	---	---
C4.5 [35]	83.60	---	---
NB [35]	87.83	---	---
BNND [35]	90.00	---	---
BNNF [35]	88.76	---	---
RBF [35]	85.00	---	---
RST+KNN [36]	81.81	---	---
RST+LSVM [36]	80.64	---	---
RST+RBF+SVM [36]	82.16	---	---
RST+DT [36]	**74.19**	**---**	**---**
RST+RF [36]	**88.66**	**---**	**---**
RST+NB [36]	**71.84**	**---**	**---**
R-HEFS+AdaBoost [37]	**84.90**	**---**	**---**
R-HEFS+SVM [37]	**85.80**	**---**	**---**
PCA+Naïve Bayes [24]	**88.38**	**---**	**---**
CfsSubsetEval+LR[24]	**86.45**	**---**	**---**
infoGainAttributeEval+J48 [24]	**83.87**	**---**	**---**
Proposed model	90.63	71	96

IV. Discussion

This work presented a new supervised learning model that enables the classification process through a decision tree in its training phase, generated from dimensional space partitions using hyperplanes. First, the model was applied to the XOR function problem, a problem that cannot be solved using a linear classifier like a simple perception; however, this problem can be solved by the proposed model. Second, the model was applied to the hepatitis B disease diagnosis problem

using the hepatitis dataset from the University of California, Irvine (UCI) Machine Learning Repository. The model was compared to other state-of-the-art models using performance evaluation metrics such as Sensitivity, Specificity, and Accuracy, yielding values of 71%, 96%, and 90.63%, respectively. Not all works shown in Table 2 evaluated Sensitivity and Specificity metrics, which is why some values are not reported. However, among those that did report both metrics, the proposed model achieved the best result in Specificity, meaning it obtained the highest percentage in detecting negative cases. On the other hand, the Sensitivity value was not the highest; other models in Table 2 showed better percentages in detecting positive cases. Nonetheless, the Accuracy value reported by the proposed model was the highest among the compared models, indicating that it achieved the best rate of correct predictions. Finally, based on the results shown in Table 2, it can be observed that the proposed model in this work is a viable option to consider among supervised learning models.

V. Conclusions and Future Work

This work presented a machine learning model that addresses the diagnosis of hepatitis B disease. The proposed model corresponds to a supervised learning model that con-structs a decision tree from the partitions of the dimensional space into sections of vectors belonging to the same class using hyperplanes. The model was applied to the hepatitis dataset from the University of California, Irvine (UCI) Machine Learning Repository. The dataset consists of 155 patient records divided into two classes (die and live) and contains 19 attributes. Approximately 48.30% of the data is missing, so K-NN imputation was used to manage the missing data, and PCA for feature selection. The model was trained using an 80-20 train-test split, with 80% corresponding to the training set and 20% to the test set. The tree generated by the model in its training phase consisted of 67 nodes and had a depth of 7. The confusion matrix generated by the model is shown in Fig. 10, and the values of Sensitivity, Specificity, and Accuracy were calculated as performance evaluation metrics. The results were compared with other works shown in Table 2. The proposed model achieved the best values for Accuracy and Specificity at 90.63% and 96%, respectively, and while it did not achieve the highest Sensitivity value, it is above the average of the models with which it was compared and that report Sensitivity. Although the proposed model shows competitive results when applied to the diagnosis of hepatitis B disease, future work could involve applying the model to other datasets and problems to compare its performance with other works in various application fields.

Conflict of Interest. No potential conflict of interest was reported by the author(s).

Acknowledgements. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, EDD, EDI, SIP, and ESCOM) and SECIHTI for their financial support in developing this work.

References

[1] S. Madihi, I. El Harrass, S. Malki, G. El Ouali Lalami, S. Nadir and M. Alaoui et al., Virological characterization of hepatitis B virus infection in Morocco: A ten-years study (2014 - 2023), *Diagnostic Microbiology and Infectious Disease*, **110** (2024), 116502. https://doi.org/10.1016/j.diagmicrobio.2024.116502

[2] R. Zhang, Z. Lin, X. Chen, Y. Zhang, Y. Xie and Y. Cai et al., Trends and multi-model prediction of hepatitis B incidence in Xiamen, *Infectious Disease Modelling*, **9** (2024), 1276 - 1288. https://doi.org/10.1016/j.idm.2024.08.001

[3] S. Yin, X. Lin, Y. Chen, S. Zheng and Z. Zhang, Spatial-temporal analysis of hepatitis B in Fujian Province, China in 2012 - 2021, *Infectious Medicine*, **3** (2024), 100110. https://doi.org/10.1016/j.imj.2024.100110

[4] T. Khan, H.A. Wahash, M.A. Abdelkawy, H.M. Al-Gannam, A.F. Al-Mutairi and A.K. Almaweri, Modeling different infectious phases of hepatitis B with generalized saturated incidence: An analysis and control, *Mathematical Biosciences and Engineering*, **21** (2024), no. 4, 5207 - 5226. https://doi.org/10.3934/mbe.2024230

[5] D. Singh, A. Kumar, P.K. Kashyap and S. Kumar, Fuzzy logic based medical diagnostic system for hepatitis B using machine learning, *Soft Computing*, (2023), 1 - 17. https://doi.org/10.1007/s00500-023-08894-3

[6] G. Adali, F.C.T. Tatli, Y.K. Ozer, H.B. Kalkan, H. Ozer and Y. Karatay et al., Chronic hepatitis B and metabolic dysfunction-associated steatotic liver disease: Metabolic risk factors are key drivers of hepatocellular carcinoma, *Heliyon*, **10** (2024), e37990. https://doi.org/10.1016/j.heliyon.2024.e37990

[7] Y. Zhou, C. Zhao and Q. Xie, How hepatitis B virus causes cirrhosis and liver cancer, *Medical Hypotheses*, **108** (2017), 52 - 53. https://doi.org/10.1016/j.mehy.2017.08.005

[8] R. Sedano, D.D. Cohen, P.C. Stock, H. Te, K.M.A. Tujios and J.R. Leef et al., Liver transplantation in acute liver failure due to hepatitis B: Two clinical cases, *Annals of Hepatology*, **21** (2021), 100107. https://doi.org/10.1016/j.aohep.2019.06.012

[9] F. Fabrizi, P.M. Mannucci and P. Messa, Hepatitis B virus vaccine and chronic kidney disease: The advances, *Nefrologia*, **41** (2021), no. 2, 115 - 122. https://doi.org/10.1016/j.nefro.2020.08.016

[10] S. Delghandi, K. Safaie, R. Mohammadian, M.R. Aghasadeghi, S.A.G. Tabatabaei and S.S.S.K. Boukany et al., An overview of occult hepatitis B infection (OBI) with emphasis on HBV vaccination, *Heliyon*, **10** (2024), e37097. https://doi.org/10.1016/j.heliyon.2024.e37097

[11] D. Villa and M.C. Navas, Vertical transmission of hepatitis B virus - an update, *Microorganisms*, **11** (2023), no. 5, 1140. https://doi.org/10.3390/microorganisms11051140

[12] A.A. Faniyi, O.K. On-Akpodia, H.J. Bassey and O. Omoruyi, Advancing public health policies to combat hepatitis B in Africa: Challenges, advances, and recommendations for meeting 2030 targets, *Journal of Medicine, Surgery, and Public Health*, **2** (2024), 100058. https://doi.org/10.1016/j.glmedi.2024.100058

[13] S. Tung-Hung and K. Jia-Horng, Role of artificial intelligence in the management of chronic hepatitis B infection, *Clinical Liver Disease*, **23** (2024), e0164. https://doi.org/10.1097/CLD.0000000000000164

[14] Y. Xiao, A.J. Thompson and J. Howell, Point-of-care tests for hepatitis B: An overview, *Cells*, **9** (2020), 2233. https://doi.org/10.3390/cells9102233

[15] V. Harabor, F.C.A.G. Sporea, R.L.R.P.D. Stan, A.I.R. Popescu and A.M.D.C. Streba, Machine learning approaches for the prediction of hepatitis B and C seropositivity, *International Journal of Environmental Research and Public Health*, **20** (2023), 2380. https://doi.org/10.3390/ijerph20032380

[16] Y. Kaya and M. Uyar, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, *Applied Soft Computing*, **13** (2013), 3429 - 3438. https://doi.org/10.1016/j.asoc.2013.03.008

[17] X. Tian, S. Li, Y. Liu, X. Yan and D. Zhu, Using machine learning algorithms to predict hepatitis B surface antigen seroclearance, *Computational and Mathematical Methods in Medicine*, **2019** (2019), 6915850. https://doi.org/10.1155/2019/6915850

[18] O. Peretz, M. Koren and O. Koren, Naive Bayes classifier - an ensemble procedure for recall and precision enrichment, *Engineering Applications of Artificial Intelligence*, **136** (2024), 108972. https://doi.org/10.1016/j.engappai.2024.108972

[19] K. Moulaei, M. Shanbehzadeh, S.S.S.A. Aghdam, H.G.H. Savoji and E.R.N. Tutti, Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis, *International Journal of Medical Informatics*, **179** (2023), 105243. https://doi.org/10.1016/j.ijmedinf.2023.105243

[20] S. Datta, S. Misra and S. Das, A feature weighted penalty-based dissimilarity measure for k-nearest neighbor classification with missing features, *Pattern Recognition Letters*, **80** (2016), 231 - 237. https://doi.org/10.1016/j.patrec.2016.06.023

[21] 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*, **108** (2012), 570 - 579. https://doi.org/10.1016/j.cmpb.2011.08.003

[22] D. Calisir and E. Dogantekin, A new intelligent hepatitis diagnosis system: PCA - LSSVM, *Expert Systems with Applications*, **38** (2011), 10705 - 10708. https://doi.org/10.1016/j.eswa.2011.01.014

[23] F. Rui, X. Yang, S. Wang, J. Wang, X. Xu and Z. Huang et al., Development of a machine learning-based model to predict hepatic inflammation in chronic hepatitis B patients with concurrent hepatic steatosis: A cohort study, *eClinicalMedicine*, **68** (2024), 102419. https://doi.org/10.1016/j.eclinm.2023.102419

[24] E.A. Bayrak, P. Kirci and T. Ensari, Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease, *International Journal of Multidisciplinary Studies and Innovative Technologies*, **3** (2019), no. 2, 135 - 138.

[25] R. Amin, I.U. Haq, F. Al-Ghamdi, A. Al-Ghamdi and K. Muhammad, Prediction of chronic liver disease patients using integrated projection-based statistical feature extraction with machine learning algorithms, *Informatics in Medicine Unlocked*, **36** (2023), 101155. https://doi.org/10.1016/j.imu.2022.101155

[26] S. Zhang, Y. Wang, W. Liu, Z. Liu, X. Gong and W. Sun, Construction of a diagnostic model for hepatitis B-related hepatocellular carcinoma using machine learning and artificial neural networks and revealing the correlation by immunoassay, *Tumour Virus Research*, **16** (2023), 200271. https://doi.org/10.1016/j.tvr.2023.200271

[27] S. Ranjbar, CART-based wide-area damping controller for inter-area oscillations in bulk power system consisting of WAMS data, *ISA Transactions*, **153** (2024), 350 - 363. https://doi.org/10.1016/j.isatra.2024.07.028

[28] M. Chen, G. Li, Z. Chen, Y. Wang, Y. Zhang and Y. Wang et al., CDRM: Causal disentangled representation learning for missing data, *Knowledge-Based Systems*, **299** (2024), 112079. https://doi.org/10.1016/j.knosys.2024.112079

[29] A. Dubey and A. Rasool, Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour, *Scientific Reports*, **11** (2021), no. 1, 24297. https://doi.org/10.1038/s41598-021-03438-x

[30] R.N. Bashir, G.M.A. Abd-Elaty, M.K. Khedher, M.A.A. El-Gawad, M.N. El-Sawy and S.O.S. Al-Duais, Principal component analysis (PCA) and feature importance-based dimension reduction for reference evapotranspiration (ET0) predictions of Taif, Saudi Arabia, *Computers and Electronics in Agriculture*, **222** (2024), 109036. https://doi.org/10.1016/j.compag.2024.109036

[31] R.K. Sachdeva, A. Negi and B.G. Prasad, A systematic method for diagnosis of hepatitis disease using machine learning, *Innovations in Systems and Software Engineering*, **19** (2023), 71 - 87. https://doi.org/10.1007/s11334-022-00509-8

[32] G. Obaido, M.A. El-Hasnony, S.A. Mostafa and M.F. Mohammed, An interpretable machine learning approach for hepatitis B diagnosis, *Applied Sciences*, **12** (2022), 11127. https://doi.org/10.3390/app122111127

[33] Y. Aminat and A. Oyelola, Hepatitis diseases prediction using machine-learning techniques, *FUDMA Journal of Sciences*, **5** (2021), no. 3, 1 - 8. https://doi.org/10.33003/fjs-2021-0503-515

[34] N. Panda, B. Pati, S.C. Satapathy and S.P. Mishra, Empirical study on different feature selection and classification algorithms for prediction of hepatitis disease, Chapter in *Technical Advances in Machine Learning for Healthcare*, Springer, Singapore, 2021, 75 - 86.

[35] L. Ozyilmaz and T. Yildirim, Artificial neural networks for diagnosis of hepatitis disease, *Proceedings of the International Joint Conference on Neural Networks*, **1** (2003), 586 - 589. https://doi.org/10.1109/IJCNN.2003.1223422

[36] K.N. Singh and J.K. Mantri, A clinical decision support system using rough set theory and machine learning for disease prediction, *Intelligent Medicine*, **4** (2024), no. 3, 200 - 208. https://doi.org/10.1016/j.imed.2023.08.002

[37] R.K. Bania and A. Halder, R-HEFS: Rough set-based heterogeneous ensemble feature selection method for medical data classification, *Artificial Intelligence in Medicine*, **114** (2021), 102049. https://doi.org/10.1016/j.artmed.2021.102049

Received: September 1, 2026; Published: September 22, 2026