Journal of Engineering Technology and Applied Sciences, 2023 *e-ISSN:* 2548-0391
Received 7 December 2022, Accepted 19 April 2023 *Vol. 8, No. 1, 15-33*
Published online 29 April 2023, Research Article *doi: 0.30931/jetas.1216025*

Citation: Koçak, M. T., Kaya, Y., Kuncan, F., "Using Artificial Intelligence Methods for Detection of HCV-Caused Diseases". Journal of Engineering Technology and Applied Sciences 8 (1) 2023 : 15-33.

USING ARTIFICIAL INTELLIGENCE METHODS FOR DETECTION OF HCV-CAUSED DISEASES

Muhammed Tayyip Koçak[a], Yılmaz Kaya[b], Fatma Kuncan[c*],

[a]*Department of Software Engineering, Faculty of Engineering and Natural Sciences, University of Istanbul Health and Technology, Turkey*
muhammed.kocak@istun.edu.tr

[b]*Department of Computer Engineering, Faculty of Engineering, Batman University, Turkey*
yilmaz.kaya@batman.edu.tr

[c*]*Department of Computer Engineering, Faculty of Engineering, Siirt University, Turkey*
*fatmakuncan@siirt.edu.tr (*corresponding author)*

Abstract

The Hepatitis C Virus (HCV) can cause chronic diseases and even lead to more serious conditions such as cirrhosis and fibrosis. Early detection of HCV infection is crucial to prevent these outcomes. However, in the early stages of infection, when symptoms are not yet evident, patients rarely undergo HCV testing. This highlights the need for alternative materials to guide HCV testing for early detection of the disease. In this study, we investigate the use of artificial intelligence technology to determine the disease status of individuals using blood data. A total of 615 individuals were included in the study. Preprocessing, filtering, feature selection, and classification processes were applied to the blood data. The correlation method was used for feature selection, where the features with high correlation values were selected and given as input to five different classification algorithms. The results of the study showed that the K-Nearest Neighbor (KNN) algorithm achieved the best classification success for detecting HCV patients, with a rate of 99.1%. This research demonstrates that artificial intelligence technology can be an effective tool for early detection of HCV-related diseases. The results indicate that the KNN algorithm can provide clear information about hepatitis infection from different blood values. Future studies can explore the use of other AI techniques and expand the sample size to improve the accuracy of the model.

Keywords: Hepatitis C virus, k-nearest neighbors, preprocessing, machine learning, classification

1. Introduction

Hepatitis C is a virus that causes many diseases that can have fatal consequences [1]. It's one of the leading causes of chronic liver disease [1]. While this virus might cause minor illnesses, it can also induce a variety of severe conditions, such as liver cirrhosis or cancer [2]. In 1975, Feinstone et al. made the initial discovery of hepatitis C. This team discovered that hepatitis A or hepatitis B viruses are typically not linked to occurrences of hepatitis associated with blood transfusions. This infection, once known as non-A-non-B hepatitis, is now recognized as hepatitis C [3], [4].

Hepatitis C virus infection affects over 58 million individuals globally, and every year, 1.5 million new cases are reported. According to the World Health Organization, approximately 290,000 people died in 2019 due to hepatitis C-related diseases, mostly cirrhosis and hepatocellular carcinoma [5].

Transmission of the HCV virus can occur for a variety of reasons. Intravenous drug usage, blood transfusions, and malpractice in the medical field are the most frequent of these. While intravenous drug use is the most frequent form of infection for developed nations, blood transfusions are the most frequent cause of transmission for developing nations [6], [7].

For a patient with hepatitis C infection, treatment may not be needed because the solution to the infection is often provided by the body's immune system. However, treatment is required if chronic hepatitis is present [8]. Pan-genotypic, direct-acting antivirals are often preferred for patients over the age of 12 in circumstances when therapy is necessary. The condition can be cured with this treatment for up to 12 weeks. However, if there is cirrhosis, this process may be prolonged [5].

Despite recent improvements in the availability of hepatitis C therapy, poor detection rates of the illness remain in undeveloped and emerging nations as a result of inadequate conditions and care. The fact that the majority of infected patients are unaware of their disease prevents them from being directed to treatment. Approximately 21% of the 58 million persons worldwide with hepatitis C infection in 2019 obtained a diagnosis of the illness, and 62% of those diagnosed received antiviral treatment, according to data from the World Health Organization [2], [5].

Based on these data from the World Health Organization, it is understood that the most important factor for the treatment of the disease is the detection of the disease. The HCV test is usually used to detect hepatitis C infection. To determine if a patient has been exposed to this virus, a blood test called the HCV test is utilized. This test, also called the HCV antibody test, measures the level of antibodies produced in the blood against hepatitis.

Due to their misconception that HCV testing is not necessary, patients with hepatitis C put off diagnosing their condition, especially in the early asymptomatic phases. Different materials are required in this circumstance for HCV testing. Developing artificial intelligence technology can be an alternative to these materials, which are necessary for the early diagnosis of the disease. Machine learning and artificial neural network from artificial intelligence methods can be used for this purpose [9].

Use of classification algorithms such as machine learning alone; It may be insufficient to solve a problem, to create a good model of the data set, and to understand whether this model

is sufficient and appropriate. For this reason, the success of the model can be increased by using supporting processes such as data filtering, preprocessing, feature selection, and feature extraction beforehand. In this work, filtering, data preprocessing, and feature selection were used before classification. As can be shown, success rates improve as a direct result of the use of these helpful resources.

For the study, the blood values of 615 individuals were pre-processed by applying normalization and filtering. Then, feature selection was performed by correlation. Finally, the features with high correlation values were given as input to different classification methods, and the results were compared. As a result of the study, the K-Nearest Neighbor Method provided the highest classification success rate with 99.1%.

2. Literature survey

Sarmento examined the inputs and targets of HCV data and made a detailed analysis of the features. Calculated the number, range of values, and averages of each attribute. It performed inference analysis across all features. A total of 78 inference analyses were carried out. Exploratory factor analysis was performed. Exploratory factor analysis is a statistical method used to describe the variability among observed, and associated variables. The purpose of doing factor analysis is to investigate some unobserved variables called factors. The correlation values (between different variables in the study) ranged from a minimum of 0.147 to 0.532. Cluster analysis was performed. Non-hierarchical clustering methods such as K-means and hierarchical ones were used, and after optimization of both methods, it was concluded that clustering of the presented data was approached by using the 3 best clusters. Finally, classification analysis was applied. The classification analysis results are shown in Table 1 [10].

Table 1. Accuracy result.

Classification methods	**Result**
AdaBoostM1	91.341
Bagging	93.379
DecisionStump	91.341
JRip	92.36
LMT	94.737
Logistic	94.228
OneR	91.511
PART	93.209
SMO	92.699
Stacking	89.304
LogitBoost	95.586
J48	92.53
IBk	92.699

Khair Ahammed et al. stated that the hepatitis C virus is a major factor in the occurrence of liver disease around the world. On the other hand, it has been explained that many methods have been developed to try to reduce the effect of the infected virus. Khair Ahammed et al. proposed a machine learning-based structure method for classifying liver disease stages in people infected with the hepatitis C virus. Individuals showing liver fibrosis disease among Egyptian patients were collected in the UCI repository. The synthetic minority oversampling technique, which increases artificial patient samples, is used to increase, and stabilize category samples. Then, different feature selection methods were applied to determine the important features of the hepatitis C virus in this data set. Various classification methods have been used to classify HCV data. After the results were analyzed, it was seen that the best method was KNN, with 94.40% accuracy. As a result, it has been useful in determining the disease caused by the hepatitis C virus [11].

According to Lailis Syafa'ah and colleagues, the principal pathogen that is transferred through the blood to people is the Hepatitis C virus. The study included 73 patients. The study looked into the classification machine learning algorithms. Their goal is to examine and assess the level of accuracy of detecting HCV illness using machine learning classification methods. To classify the data, neural networks, decision trees, PSO, GA, logistic regression, and support vector machines were utilized. When the results are compared with the accuracy rates of KNN, Naive Bayes, and RF methods, which are 89.43%, 90.24%, and 94.31%, respectively, it is seen that the accuracy rate of the NN technique is higher at 95.12% [12].

In this study, John Rigg et al. aim to provide evidence for the application of machine learning on electronic medical record data to prioritize patients for hepatitis C virus testing to improve resource utilization efficiency, reduce physician burden, and save healthcare costs. After downsampling unlabeled patients to help the algorithm learn, 16.2 million unlabeled patients were included in the analysis. The algorithm's accuracy was 2%, 0.4%, and 0.12% at 5%, 20%, and 50% recall, while its specificity was 99.9%, 99.0%, and 90%, respectively. The AUCROC value was 0.81 [13].

Hepatitis C (HCV) is a micro-infectious infection that causes inflammation of the liver, which can be fatal. It was explained that in every medical treatment, it is very important to determine the exact treatment response in order to reduce the symptoms of the disease. In this research, Utkrisht Singh et al. used a two-dataset technique to detect the hepatitis C virus in the general population. Many machine learning methods were applied to the classification dataset. The best result on the test set with these various machine learning algorithms was a logistic regression score of 0.9430 [14].

3. Material and method

Within the scope of the study, on the data set containing the blood values of 615 individuals; HCV infection detection was performed using an artificial intelligence algorithm consisting of data preprocessing, filtering, feature extraction, and classification processes.

3.1 Data set

In the study, an online-accessible data set, in which blood values of blood donors and hepatitis patients were recorded, was used [15].

In this data set, a total of 615 individuals, 238 women and 377 men, aged between 19 and 77, are identified by age, gender, disease status, and 10 different blood characteristics. All the characteristics of the individuals except their disease status and gender were presented as numerical data. Table 2 shows the inputs and outputs in the data set.

Table 2. Data set inputs and output columns

Age	Gender	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Disease

In the data set, the disease status of individuals is specified as five different classes to be used as output data. Their disease status and the number of people belonging to these classes are indicated in Figure 1. The number of healthy people has the highest number, at 533. The number of people with cirrhosis, fibrosis, and hepatitis is 30, 21, and 24, respectively. There is a class of suspects with the least number of people, with 7 people.

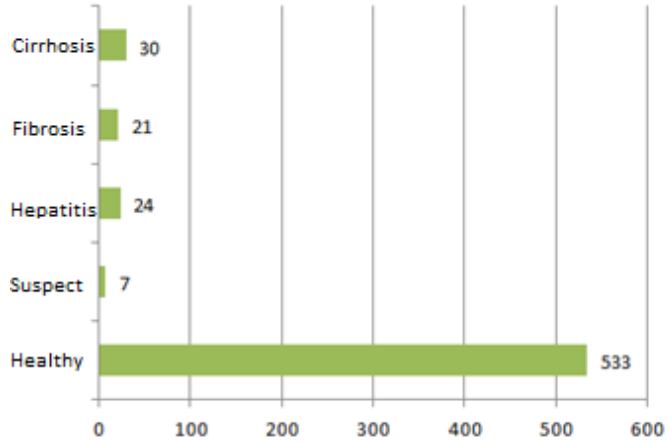

Figure 1. Disease conditions and number of persons

There are a total of 615 people in the data set, including 377 men and 238 women. It is seen that the number of male people is 139 more than the number of female people. Figure 2 shows the gender distribution graph.

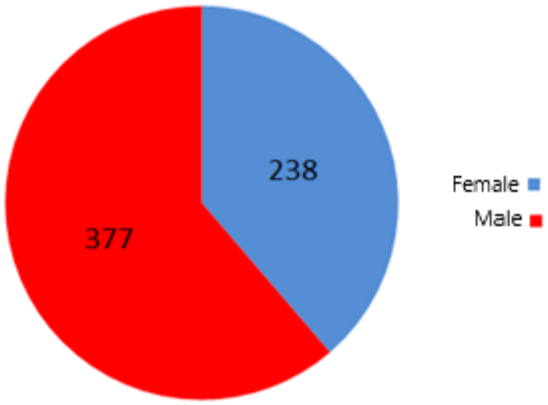

Figure 2. Gender distribution

The number of male persons comprises 61.3% of all numbers.

In Figure 3, a graph of the age histograms of the individuals is given. When Figure 3 is examined, it is seen that the majority of individuals are in middle age. The age of the people in the dataset is between 19 and 77 years. The average age of the people in the dataset is 47.4081. 46-year-olds are in the majority.

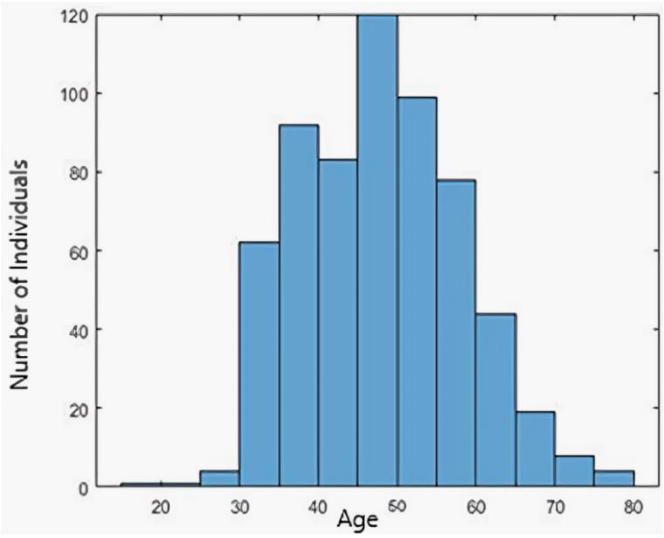

Figure 3. Age histograms of individuals

3.1.1 Albumin (ALB)

The albumin blood test measures the amount of albumin in the blood. The ideal value for ALB is in the range of 34 to 54. Figure 4 shows a chart of the histograms of ALB values in the dataset. The amount of albumin in the blood can give information about whether there are liver or kidney diseases. High levels can be a sign of dehydration. As can be seen in Figure 4, the majority of the frequency range of the values is at normal values.

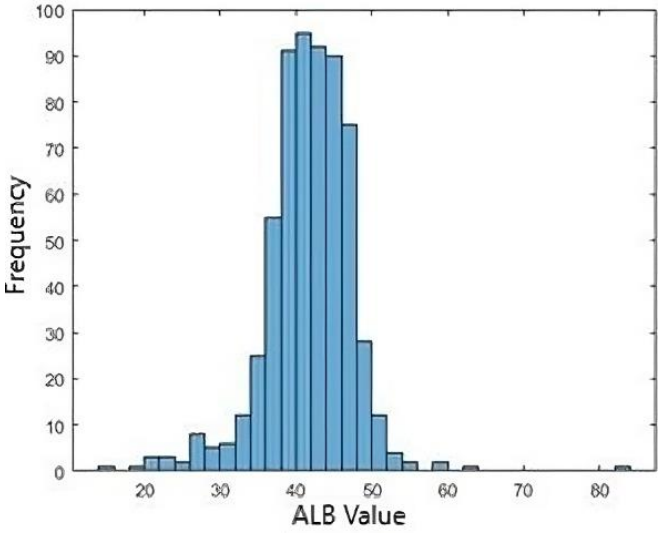

Figure 4. ALB value histogram chart

3.1.2 Alkaline phosphatase (ALP)

Alkaline phosphatase (ALP) is a protein found in all body tissues. The ideal value for ALP is between 44 and 147. In Figure 5, the histogram graph of the ALP values in the data set is given. Tissues with higher amounts of ALP include scatters, bile ducts, and bones. A blood test can be done for ALP unit.

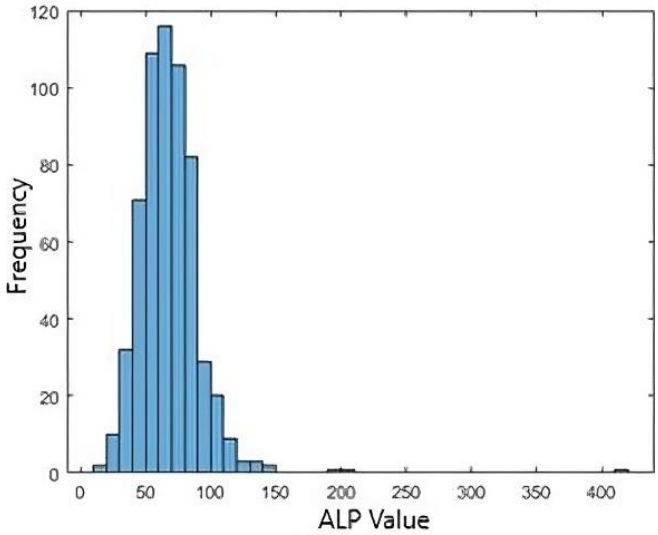

Figure 5. ALP value histogram chart

3.1.3 Alanine aminotransferase (ALT)

The alanine aminotransferase (ALT) test is a blood test that checks for liver damage. This value can be looked at to understand the damage caused by the disease or a drug to the liver. The ideal value for ALT is between 7 and 56. Figure 6 shows the graph of the histograms of the ALT values in the data set. It is seen that the normal range is largely in agreement with the histogram graph in Figure 6.

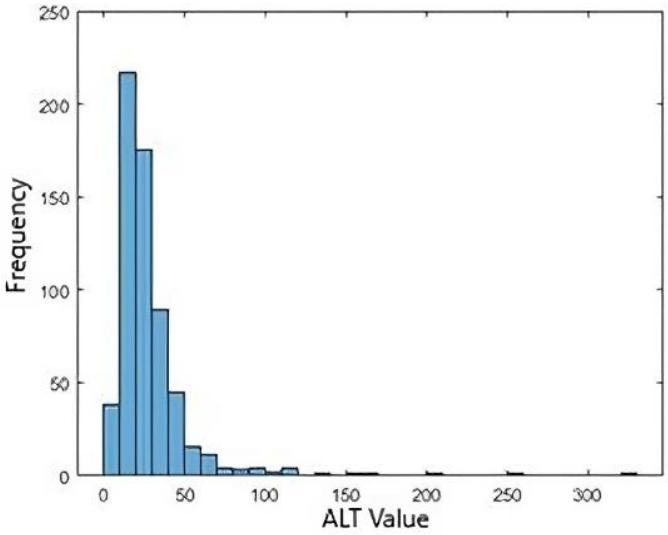

Figure 6. ALT value histogram chart

3.1.4 Aspartat aminotransferase (AST)

The aspartate aminotransferase (AST) blood test measures the level of the AST enzyme in the blood. The ideal value for AST is between 5 and 40. In Figure 7, the graph of the histograms of the AST values in the data set is given. It is seen in the light of Figure 7 and ideal values that the AST values in the data set are significant.

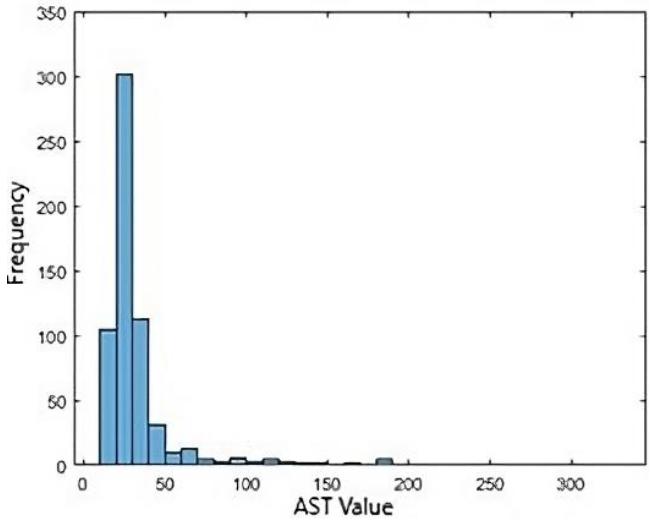

Figure 7. AST value histogram chart

3.1.5 Bilirubin (BIL)

Bilirubin is a yellowish pigment made during the normal breakdown of red blood cells. BIL measures the total level of bilirubin in the blood, and the ideal value for BIL is in the range of 0.3 to 1. In Figure 8, the graph of the histograms of the BIL values in the data set is given. In the data set, it is seen that these values are close to zero, as in normal values.

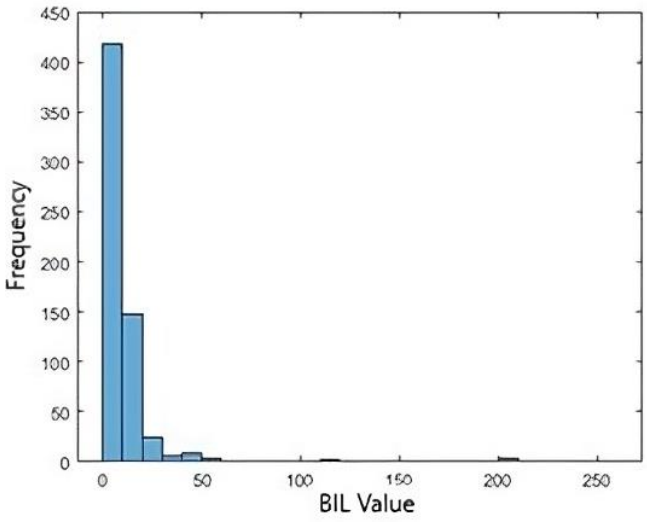

Figure 8. BIL value histogram chart

3.1.6 Serum cholinesterase (CHE)

Serum cholinesterase is a blood test that looks at the levels of two substances that help the nervous system function properly. These are called acetylcholinesterase and pseudocholinesterase. The ideal value for CHE is between 8 and 18. In Figure 9, the graph of the histograms of the CHE values in the data set is given.

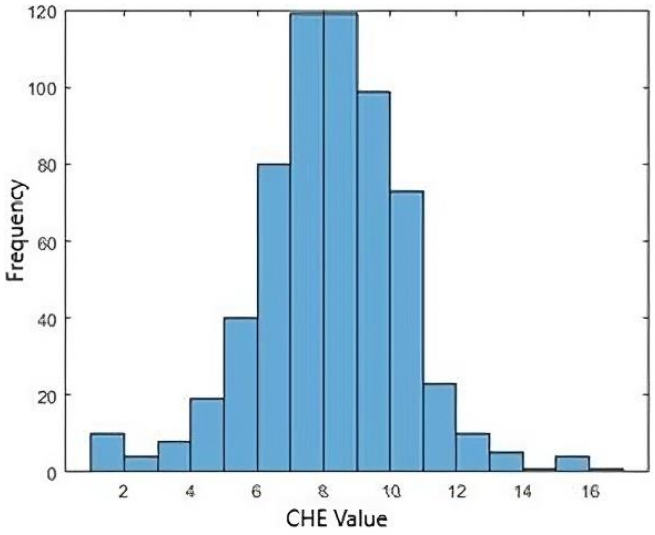

Figure 9. CHE value histogram chart

3.1.7 Lipid profile (CHOL)

Also called a "lipid panel" or "lipid profile," this test is a blood test that can measure the amount of cholesterol and triglycerides in the blood. The ideal value for CHOL is in the range of 5.2 to 3.4. In Figure 10, the graph of the histograms of the CHOL values in the data set is given.

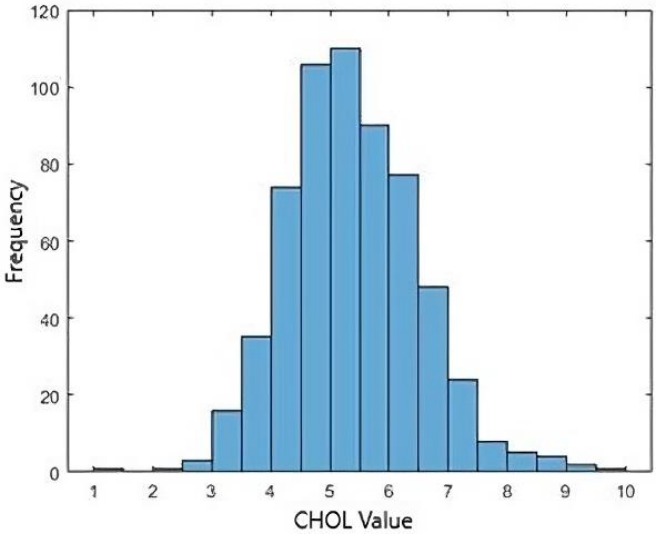

Figure 10. CHOL value histogram chart

3.1.8 Creatinine (CREA)

Creatinine is a measure of how well the kidneys are doing their job of filtering waste out of the blood. The ideal value for creatinine is between 61.9 and 114.9 for men and 53 and 97.2 for women. Women generally have lower creatinine levels than men. In Figure 11, the graph of the histograms of the CREA values in the data set is given.

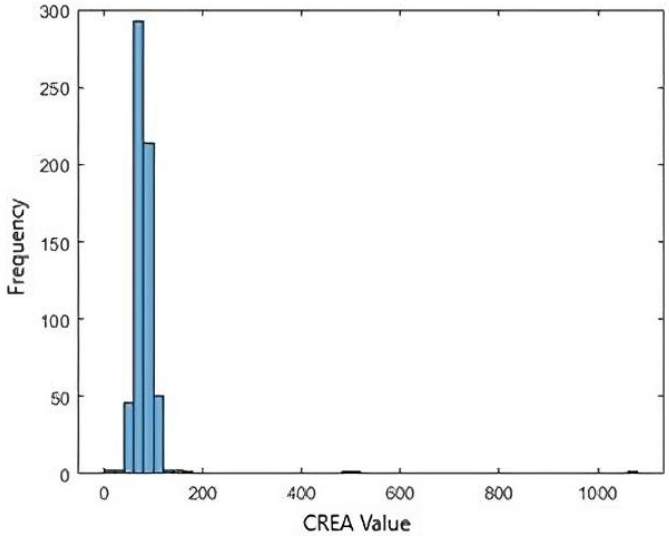

Figure 11. CREA value histogram chart

3.1.9 Gamma glutamyl transferase (GGT)

The GGT test measures the level of an enzyme called GGT, which is found at high levels in the liver, kidney, pancreas, heart, and brain. The ideal GGT value is in the range of 5 to 40. In Figure 12, the graph of the histograms of the GGT values in the data set is given.

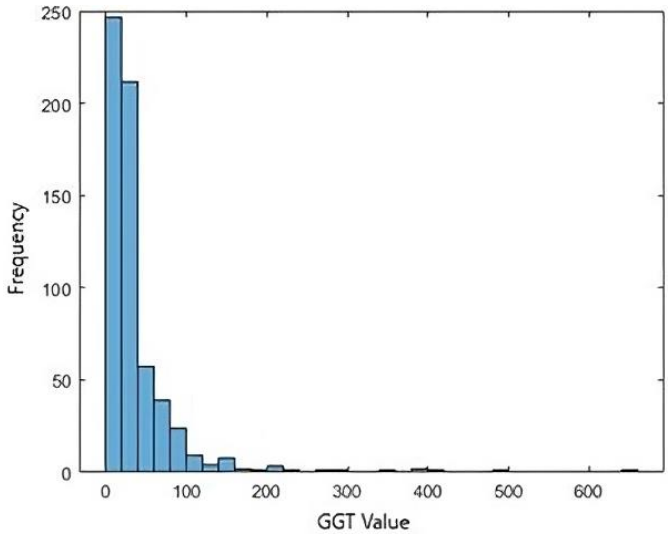

Figure 12. GGT value histogram chart

3.1.10 Total protein (PROT)

The total protein test measures the total amount of albumin and globulin proteins found in the liquid portion of blood. The ideal value for PROT is between 60 and 83. In Figure 13, the graph of the histograms of the PROT values in the data set is given.

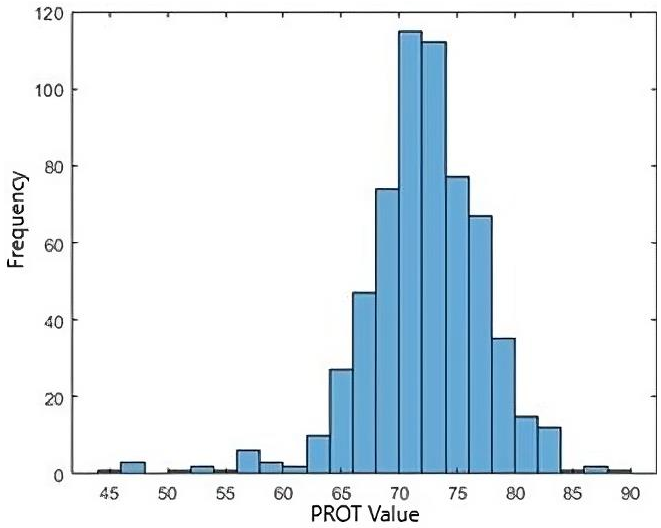

Figure 13. PROT value histogram chart

3.2 Data preprocessing and filtering

Some processes have been performed in order to make the data presented in the data set suitable for classification. First, data with missing blood information was selected and eliminated. Then, disease status and gender information, which were not given as numerical values, were digitized.

Gender information is specified in the data set and digitized as 0 for male and 1 for female. The disease status information to be given as an output to the classification process is 1 for healthy, 2 for hepatitis, 3 for fibrosis, 4 for cirrhosis, and 5 for suspicious conditions.

The performance of artificial neural networks and machine learning methods as classifier algorithms has been evaluated. In order to be able to classify with the artificial neural network, in addition to the data pre-processing part, the output information is converted to binary codes in order to observe the correct classification success (healthy: 10000, hepatitis: 01000, fibrosis: 00100, cirrhosis: 00010, suspect: 00001).

A moving average filter is used to de-noise the data. The moving average filter smoothest the data by replacing each data point with the average of neighboring data points defined within the range. This operation is part of the low-pass filter's logic. Thanks to this filter, the data is purified from noisy input and softened.

3.3 Feature selection

Feature selections are used to make the data more meaningful. In this study, feature selection was made by using the correlation technique. Correlation is often used to explain how a linear relationship exists between variables. In this way, highly correlated features can be detected. Low-correlated features should be ignored as they will have a low correlation. Also, highly correlated features are more linearly dependent and, therefore, have almost the same effect on the outcome. Therefore, when two features are highly correlated, one of the two features can be dropped [16]. In this study, correlations among features were determined, and feature selection was performed according to these results.

3.4 Artificial neural networks

The method of artificial neural networks (ANN) was used to classify the prepared data. Artificial neural networks are an artificial intelligence method inspired by the functioning of neurons in the human brain [17].

In the algorithm, neurons are connected to each other by features, and each link has a numerical weight that expresses the importance of the input data. Weights are the main tool of long-term memory in neural networks. The neural network operates by repeatedly adjusting the weights to perform learning [18].

3.5 Machine learning

3.5.1 Decision trees

Decision Tree is an algorithm that can be used for both classification and regression problems, but it is mostly a supervised machine learning technique that is preferred to solve classification problems. A decision tree is a tree structure similar to a flowchart where each internal node represents a test on an attribute, each branch reflects the result of the test, and each leaf node provides its class label (or end node) [19].

A decision node and a leaf node are two nodes of a decision tree. Leaf nodes are the result of these decisions and have no additional branches, whereas decision nodes are used to make any decision and have many branches. It is easy to understand as it follows the same process that a person follows when making any decision in real life. However, decision trees become complex in applications where they are multi-layered. As a result, a decision tree simply asks a question, divides the tree into subtrees, and classifies according to the answer (yes/no) [20].

3.5.2 K-nearest neighborhood (KNN)

The KNN classification method is one of the most straightforward. It is a method frequently used in supervised learning, regression problems, and classifications [21, 22]. It is generally characterized as a lazy algorithm, although it achieves high success rates. It is not recommended for use with large datasets. In large datasets, the cost of calculating the distance between the new point and every existing point is very high, causing the performance of the algorithm to degrade.

The purpose of the KNN classification algorithm is to search for points closest to the new point. K allows specifying how many nearest neighboring points there are to a point of

unknown class to be referenced. Usually, the number K is taken as an odd number. The test sample is determined to belong to this category if the data belonging to that category is greater than the amount of K data received [21].

To determine which data points are closest to a particular query point, the distance between the query point and other data points will need to be calculated. These distance measurements help establish decision boundaries that divide query points into different regions. Euclidean and Manhattan distance criteria are used for distance calculations when determining the nearest neighbors ([1], [2] and, [3]).

$$Euclidean(X,Y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (1)$$

$$Manhattan(X,Y) = \sum_{i=1}^{k}|x_i - y_i| \quad (2)$$

$$Minkowski(X,Y) = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \quad (3)$$

To summarize, the purpose of the k-nearest neighbor algorithm is to determine the nearest neighbors of a given query point so that a class label can be assigned to that point.

3.5.3 Support vector machines (SVM)

Support Vector Machines is a supervised machine learning algorithm that can be used for classification, pattern recognition, and regression. However, it is generally used in classification analysis.

This algorithm performs the classification by taking the entered training data and finding the hyperplane line that defines the boundary between the classes. With this method, after the hyper-plane is detected, the class of the entered test data is included in the region of the border, making classification possible [21].

The algorithm of the Support Vector Machines method can be formulated as equation (4).

$$\gamma = \begin{cases} 0 if w^T.x + b < 0, \\ 1 if w^T.x + b \geq 0 \end{cases} \quad (4)$$

3.6 Performance metrics

In order to evaluate the performance of the classification techniques used for HCV data classification, several metrics such as accuracy, precision, F-score, and sensitivity were employed. These metrics were calculated using the parameters obtained from the Confusion Matrix presented in Table 3.

Table 3. Confusion matrix

True values			
Prediction values		**Positive**	**Negative**
	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Specified in the table; T, F, P, N letters respectively; represents true, false, positive, and negative. As an example, TP represents the number of correctly classified positive data.

Accuracy 3.6.1 The number of correctly classified data. Accuracy indicates whether the model is generally correct. The equation used in the calculation is given as (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision 3.6.2 The ratio of positive-correct predictions to total positive predictions. Equation (6) used in the calculation is given.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Sensitivity 3.6.3 It is expressed as the ratio of correctly identified positive data (TP) to the number of truly positive data (TP+FN). The equation used in the calculation is given as (7).

$$Sensitivity: = \frac{TP}{TP+FN} \quad (7)$$

F-Score 3.6.4 This criterion is the harmonic mean of the precision and sensitivity measures. It is calculated by the following equation (8).

$$F - Score: = \frac{2\times TP}{2\times TP+FP+FN} \quad (8)$$

4. Result

In this study, machine learning and artificial neural network methods were implemented using the MATLAB program to classify hepatitis disease based on the blood values of various patients. The input data used in the dataset included 10 different blood values, gender information, and age information. The output data comprised of hepatitis disease states. The data was divided into 80% for training and 20% for testing, and the classification process was applied. The comparison of results revealed that the algorithm with the highest classification success rate of 99.1% was achieved by applying a combination of techniques including the deletion of missing data, data filtering, normalization, feature selection, and classification using the K-Nearest Neighbor method, with a K-value of 3. Additionally, an artificial neural network with 1 fully connected layer and 3 hidden layers of 25, 10, and 10 neurons was used to classify hepatitis disease based on blood values, with a ReLU activation function and an iteration limit of 1000.

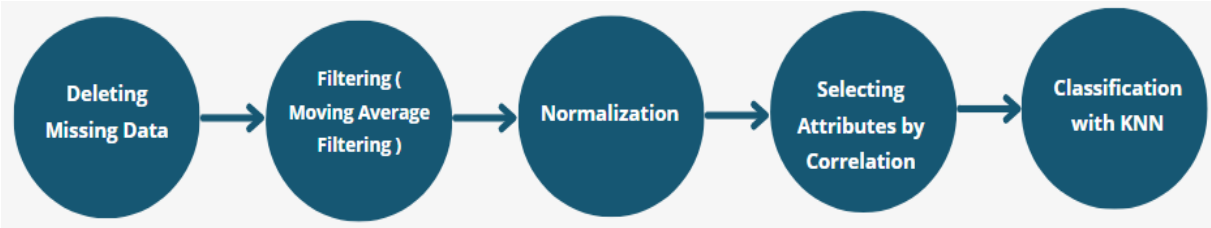

Figure 14. Algorithm diagram

The flowchart of the algorithm that yielded the highest performance can be seen in Figure 14. Additionally, the Confusion Matrix resulting from the application of data preprocessing techniques such as deletion of missing data, filtering, normalization, feature selection and the KNN classification method is presented in Table 4.

Table 4. Confusion matrix occurred by KNN classification.

		Prediction values				
		Healthy	**Hepatitis**	**Fibrosis**	**Cirrhosis**	**Suspect**
True values	**Healthy**	105	0	0	0	0
	Hepatitis	0	5	0	0	0
	Fibrosis	0	0	2	0	0
	Cirrhosis	1	0	0	3	0
	Suspect	0	0	0	0	1

Table 4 illustrates that the majority of the individuals in the dataset are in a healthy state. However, it is crucial to accurately classify the minority data, as they represent the patients with hepatitis infection. To achieve this, the success criteria were calculated using the macro average when analyzing the confusion matrix. These success criteria are presented in Table 5.

Table 5. Success criteria

Accuracy	0,99057
Precision	0,96
Sensitivity	0.93333
Specificity	0.99804
F-Score	0.93778

In the algorithm where 99.1% of the results are obtained with the K-Nearest Neighbor (KNN) method, with k as 3, normalization and deletion of missing data were applied as data preprocessing. After these processes, filtering and feature selection were applied. By keeping the same operations constant, the results of different classification algorithms were evaluated and recorded in Table 6.

Table 6. Classification accuracy results

Classification methods	Result
Decision tree	94.9
K-Nearest neighborhood	99,1
Support vector machine	97.4
Ensemble classification	96.6
Artificial neural networks	94.2

To gain insights into the impact of data filtering and feature selection on the classification results, the success rates of various preprocessing techniques were compared and recorded in Table 7. This was achieved by applying direct classification algorithms to digitized data in its raw form.

Table 7. Classification success results without feature selection and filtering

Classification methods	Result
Decision tree	91.1
K-Nearest neighborhood	90.2
Support vector machine	91.9
Ensemble classification	92.7
Artificial neural networks	89.1

The effects of applying filtering and deletion techniques on missing data in the algorithm with the highest value were evaluated by comparing the classification results obtained through various data preprocessing alternatives, as presented in Table 8.

Table 8. Comparison of classification results

		Filtering	
		By applying	**Not being applied**
Missing data	**By deleting**	KNN %99,1	SVM %96,6
	Predicting	SVM %96,7	ANN %94,2

The results presented in Table 8 demonstrate the benefits of applying filtering techniques to blood value data when using artificial intelligence methods. By filtering the data, the relationships between blood values become more apparent, resulting in more accurate classifications. Furthermore, this study has shown that deleting missing or incomplete data points is more effective than using imputation techniques to fill in the missing values. This approach allows for more accurate and reliable results.

5. Discussion

The results of this study indicate that using the KNN classification method with data preprocessing, filtering, normalization, and feature selection can be an effective tool for classifying HCV patients. The use of artificial intelligence methods, such as KNN, can provide valuable insights for early detection and diagnosis of hepatitis infection. However, to improve the success rate of the classification, the dataset should be expanded to include more data for all disease states. This is especially important since the study showed that information about hepatitis infection can be obtained through different blood values.

The artificial intelligence method used in this study can analyze blood information taken for a different purpose from a symptomless hepatitis patient who has no knowledge about the disease, thus increasing the chances of early detection and diagnosis. The application of this method in the medical field can be of critical value for early detection of hepatitis infection and initiation of treatment before the disease becomes chronic.

In future studies, artificial intelligence can be used for the early diagnosis of different diseases. With larger and more meaningful data, automatic detection of HCV and many more diseases can be achieved. In this way, diseases can be diagnosed early, and measures taken against them. Furthermore, the dataset utilized in this study can be enlarged to include additional samples from other areas and ethnic groups. Additionally, several machine learning algorithms can be assessed to determine the best approach for categorizing HCV patients. While our results demonstrate the effectiveness of our proposed method in classifying hepatitis disease based on blood values, it is important to acknowledge that these results are based on hard certainty. In future works, we plan to extend our research by applying statistical uncertainty and hypothesis testing to further validate our findings and improve the confidence in our results. In conclusion, the results of this study have demonstrated the effectiveness of machine learning and artificial neural network methods in classifying hepatitis disease according to blood values of different patients. While the algorithm used in this study achieved a high classification success rate of 99.1%, it is acknowledged that the use of a single split of 80/20% for validation may not be sufficient for such a sensitive case. In future studies, it would be beneficial to explore the use of cross-validation techniques to further improve the accuracy of the results. Additionally, the use of logarithm of certain features may also yield more accurate results. Overall, this study provides a valuable foundation for future research in this field.

Incorporating additional medical data into the study, such as demographic information and lifestyle behaviors, can give a more holistic perspective of the condition and enhance classification accuracy. Furthermore, the generated model should be evaluated on a larger patient population in a clinical setting to confirm its performance and utility in real-world circumstances. Finally, this research should be expanded to include other forms of hepatitis and other illnesses, improving diagnostic efficiency and accuracy.

Conflict of interest

No conflict of interest was declared by the authors.

Acknowledgement

This study was carried out in Siirt University Engineering Faculty Human-Computer Interaction Laboratory. The authors of this article thank the Human-Computer Interaction Laboratory staff for their support.

References

[1] CDC, "Viral Hepatitis", https://www.cdc.gov/hepatitis/hcv/index.htm, (2020).

[2] ECDC, "Hepatitis C", https://www.ecdc.europa.eu/en/hepatitis-c, (2022).

[3] Durmuş, M. E., "Buz Dağının Görünen Kısmı: Hcv Pozitif Hastalarda Tedaviye Ulaşma Oranları, Hekimlerin Yaklaşım Ve Farkındalıklarının Değerlendirilmesi", T.C. Sağlık Bilimleri Üniversitesi Antalya Sağlık Uygulama ve Araştırma Merkezi, Antalya, (2022).

[4] Feinstone, S.M., Kapikian, A.Z., Purcell, R.H., Alter, H.J., Holland, P.V., "Transfusion-Associated Hepatitis Not Due to Viral Hepatitis Type A or B", New England Journal of Medicine, vol. 292, no. 15, pp. 767–770, 1975, doi: 10.1056/NEJM197504102921502.

[5] WHO, "Hepatitis C." Jun. 2022.

[6] Dumanoğlu, B., "İstanbul medeniyet üniversitesi Göztepe Eğitim ve Araştırma Hastanesi`nde 2016-2018 yılları arasında direkt etkili antiviral tedavi alan kronik hepatit C hastalarının klinik, laboratuvar ve demografik özelliklerinin retrospektif incelenmesi", https://acikbilim.yok.gov.tr/handle/20.500.12812/290084, (2019).

[7] Maheshwari, A., Thuluvath, P. J., "Management of acute hepatitis C", Clin Liver Dis, 14(1) (2010) : 169-176.

[8] Strader, D. B., Wright, T., Thomas, D. L., Seeff, L. B., "Diagnosis, management, and treatment of hepatitis C", Hepatology 39(4) (2004) : 1147-1171.

[9] Demir, N., Kuncan, M., Kaya, Y., Kuncan, F., "Multi-Layer Co-Occurrence Matrices for Person Identification from ECG Signals.", Traitement du Signal 39(2) (2022).

[10] Sarmento, R., "Hepatitis C Records - A Complete Statistical Analysis.", Jan. 2021. doi: 10.13140/RG.2.2.22345.16481.

[11] Ahammed, K., Satu, M.S., Khan, M.I., Whaiduzzaman, M., "Predicting infectious state of hepatitis C virus affected patient's applying machine learning methods", in 2020 IEEE Region 10 Symposium (TENSYMP), (2020) : 1371-1374.

[12] Syafa'ah, L., Zulfatman, Z., Pakaya, I., Lestandy, M., "Comparison of Machine Learning Classification Methods in Hepatitis C Virus", Jurnal Online Informatika 6(1) (2021) : 73, Jun., doi: 10.15575/join.v6i1.719.

[13] Rigg, J., Doyle, O., McDonogh, N., Leavitt, N., Ali, R., Son, A., Kreter, B., "Finding undiagnosed patients with hepatitis C virus: an application of machine learning to US ambulatory electronic medical records", BMJ Health Care Inform. 30(1) (2023) doi: 10.1136/bmjhci-2022-100651.

[14] Singh, U., Gourisaria, M.K., Mishra, B.K., "A Dual Dataset approach for the diagnosis of Hepatitis C Virus using Machine Learning", in 2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022, 2022. doi: 10.1109/CONECCT55679.2022.9865758.

[15] Lichtinghagen, Ralf., Klawonn, F., Hoffmann, G., "UCI Machine Learning Repository", Available: http://archive.ics.uci.edu/ml/datasets/HCV+data, (2020).
[16] Hall, M.A., "Correlation-based feature selection for machine learning", The University of Waikato, (1999).
[17] Freeman, J.A., Skapura, D.M., "Neural networks: algorithms, applications, and programming techniques.", Addison Wesley Longman Publishing Co., Inc., (1991).
[18] Negnevitsky, M., "Artificial intelligence: a guide to intelligent systems.", Pearson education, (2005).
[19] Kim, H., Koehler, G.J., "Theory and practice of decision tree induction", Omega (Westport) 23(6) (1995) : 637-652.
[20] Suthaharan, S., Suthaharan, S., "Decision tree learning", Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning (2016) : 237–269.
[21] Altman, N.S., "An introduction to kernel and nearest-neighbor nonparametric regression", Am Stat 46(3) (1992) : 175-185,.
[22] Cover, H., Hart, P., "Nearest neighbor pattern classification", IEEE Trans. Inf. Theory 13(1) (1953) : 21-27.