RESEARCH ARTICLE OPEN ACCESS

Manuscript received February 10, 2024; revised March 18, 2024; accepted March 20, 2024; date of publication March 27, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i2.384

How to cite: Mahmud Mahmud, Irwan Budiman, Fatma İndriani, Dwi Kartini, Mohammad Reza Faisal, Hasri Akbar Awal Rozaq, Oktay Yildiz, and Wahyu Caesarendra. Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 2, pp. 116-124, April 2024.

Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease

Mahmud Mahmud[1], Irwan Budiman[1,*], Fatma Indriani[1], Dwi Kartini[1], Mohammad Reza Faisal[1], Hasri Akbar Awal Rozaq[2], Oktay Yildiz[3], and Wahyu Caesarendra[4]
[1] Faculty of Computer Science, Lambung Mangkurat University, South Kalimantan, Indonesia
[2] Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara, Turkey
[3] Faculty of Engineering, Department of Computer Engineering, Gazi University, Ankara, Turkey
[4] Faculty of Integrated Technologies, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam
*Corresponding author: Irwan Budiman (e-mail: irwan.budiman@ulm.ac.id).

Acknowledgment: Lambung Mangkurat University and Gazi University supported this work by providing valuable resources and support.

ABSTRACT Hepatitis C, a significant global health challenge, affects 71 million people worldwide, with severe complications such as cirrhosis and hepatocellular carcinoma. Despite its prevalence and availability in rapid diagnostic tests (RDTs), the need for accurate early detection methods remains critical. This research aims to enhance hepatitis C virus classification accuracy by integrating the C5.0 algorithm with Chi-Square feature selection, addressing the limitations of current diagnostic approaches and potentially reducing diagnostic errors. This research explores the development of a machine learning model for hepatitis C prediction, utilizing a publicly available dataset from Kaggle. It encompasses preprocessing techniques such as label encoding, handling missing values, normalization, feature selection, model development, and evaluation to ensure the model's efficacy and accuracy in diagnosing hepatitis C. The findings of this study reveal that implementing Chi-Square feature selection significantly enhances the effectiveness of machine learning algorithms. Specifically, the combination of the C5.0 algorithm and Chi-Square feature selection yielded a remarkable accuracy of 96.75%, surpassing previous research benchmarks. This highlights the potent synergy between advanced feature selection techniques and machine learning algorithms in improving diagnostic precision. The study conclusively demonstrates that machine learning is an effective tool for detecting hepatitis C, showcasing the potential to enhance diagnostic accuracy significantly. As a future recommendation, adopting AutoML is suggested to periodically automate the selection of the optimal algorithm, promising further improvements in detection capabilities.

INDEX TERMS C5.0 Algorithm, Feature Selection, Hepatitis C Disease, Machine learning

I. INTRODUCTION

Hepatitis is a disease that attacks the human body organ, namely the liver [1]. Hepatitis is divided into three types: A, B, and C. However, one of the three types of disease that is very dangerous is hepatitis C because this type is considered "the silent killer" [2]. According to data from the World Health Organization (WHO), in 2021, it was shown that as many as 1% or 71 million people worldwide were infected with the hepatitis C virus (HCV), of which 399 thousand people died from hepatitis C, mainly due to cirrhosis and hepatocellular carcinoma (primary liver cancer) [3]. Hepatitis C disease, if not treated quickly, can persist and develop into chronic hepatitis C. Some complications can occur due to hepatitis C infection. Some complications that can occur due to hepatitis C infection are liver cirrhosis and liver cell carcinoma [4]. Seeing the data of sufferers and the impact of hepatitis disease, it is necessary to take care to inhibit the development of hepatitis C disease. One of the efforts that can be made is to conduct screening to detect hepatitis C disease.

According to Sachdeva et al, in today's health world, medical records store the symptoms and diagnosis of a patient's illness. It can be helpful for health experts to aid in making decisions on the diagnosis of the patient's disease [5]. In addition, medical record data can also be utilized to detect diseases early in the technology field. One of the ways that can be done for early detection of disease, especially hepatitis C disease, is through essential health services such as First Level Health Facilities (FKTP) or Puskesmas by using Rapid Diagnosis Test (RDT) [6]. RDT is a medical diagnostic method designed to provide results quickly and easily, usually within minutes, without the need for sophisticated laboratory equipment or expertise to operate it.

RDTs are commonly used in various health fields, including diagnosing infectious diseases, such as malaria, HIV, hepatitis, dengue fever, and so on. These methods are often used in remote areas or developing countries where access to laboratory facilities may be limited [7]. The working principle of RDTs varies depending on the type of disease being diagnosed. However, generally, RDTs use immunochemical or biochemical techniques to detect the presence of antigens or antibodies associated with a particular disease in biological samples, such as blood, urine, or saliva [8]. Although RDTs have advantages in speed and ease of use, they often have lower sensitivity and specificity compared to more sophisticated laboratory diagnostic methods. Therefore, RDT results usually require further confirmation using more accurate diagnostic procedures [9]. In addition to RDTs, detection can also be done with the help of artificial intelligence technologies, such as machine learning.

Machine learning (ML) is a subfield of artificial intelligence in which computers, often called machines, are programmed to perform tasks automatically [10]. Machine learning combines mathematical models and sophisticated algorithms to carry out its functions. It plays an essential role in the hepatitis C diagnosis process, mainly due to its ability to manage big data and recognize patterns [11]. With the machine learning technology currently developing and the patient's medical record data, it should be a solution to help make decisions on the diagnosis of patient diseases. The utilization of machine learning technology using this classification method can be used as an initial detection of whether the patient tends to develop hepatitis C disease or not.

Several previous studies have demonstrated the effectiveness of machine learning in diagnosing hepatitis C. According to Yağanoğlu [12], using machine learning algorithms can help diagnose patients affected by chronic hepatitis C quickly. In another study conducted by Butt et al. [13] stated that machine learning is very effective in detecting hepatitis C disease in patients, with an accuracy rate of 98% in predicting hepatitis C disease. Another study also conducted by Akella and Akella [14] stated that machine learning is becoming an interesting analytical tool along with the progress of modern preventive care. From these three studies, it can be concluded that machine learning is essential in diagnosing hepatitis C disease. One of the algorithms used in machine learning is the Decision Tree 5.0 algorithm.

One of the algorithms used in machine learning is the Decision Tree 5.0 algorithm. The C5.0 algorithm is a data classification algorithm. C5.0 is an improvement of the previous algorithms created by Ross Quinlan in 1987, namely ID3 and C4.5 [15]. Compared to other classifiers, C5.0 is a classifier that classifies data in less time to produce a decision tree that requires the least amount of memory and increases accuracy [16]. According to Dalal et al. [17], in their research, using the C5.0 algorithm to classify coronary artery disease resulted in an accuracy of 85%. From previous research, it can be concluded that the C5.0 algorithm is accurate.

Feature selection is also essential in machine learning when it involves attributes in data with high dimensionality and noise [18]. One commonly used feature selection method is Chi-Square, which helps select relevant features in a dataset [19]. The Chi-Square method measures the relationship between category attributes in a dataset and category target variables [20]. Robinson Spencer et al. [21], in their research, used the BayesNet algorithm with an increase in chi-square accuracy of 5%. In another study conducted by Rosidin et al. [22] in their research using the Naïve Bayes algorithm with chi-square for the classification Covid-19 Data, the accuracy increased by 2.23%. From previous research, it can be concluded that the algorithm combined with chi-square can improve accuracy.

Based on the exposure of these problems, this research will detect the classification of the hepatitis C virus using the C5.0 algorithm combined with chi-square in feature selection. This research is expected to produce high accuracy to reduce the risk of errors in detecting the hepatitis C virus. Matters related to this research will be explained in the following chapters, namely related works, methods and materials used, research results, and conclusions reached.

II. MATERIALS AND METHODS

The research uses one of the machine learning techniques, especially supervised learning, namely the classification method with the C5.0 decision tree algorithm. In this research, several stages can be seen in the following research flowchart.

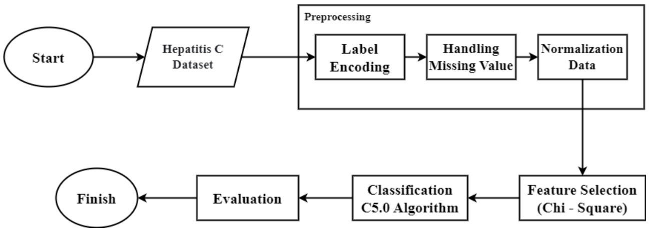

FIGURE 1. Research Flowchart Model

In FIGURE 1, the first process is data collection. After that, the data that has been obtained will enter the data pre-processing stage, which consists of handling missing values, label encoding, and normalization. Then, the next step is feature selection. Then, the next step is modeling using the C5.0 algorithm and evaluation calculations. A complete explanation will be given in the following sub-chapter.

A. DATA COLLECTION

The research data in this study was taken from the Kaggle website. This dataset follows the link https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset [23]. This data consists of 13 features and 615 rows of data containing medical records from patients who are detected to have hepatitis C disease and patients who are detected to be healthy. For each attribute of the data can be seen in the following table.

TABLE 1
Description dataset attributes

Features	Data Type	Description
Category	Binary (0,1)	Label
Age	Numerical	Attribute
Sex	Binary (0,1)	Attribute
Albumin Blood (ALB)	Numerical	Attribute
Alkaline Phosphatase (ALP)	Numerical	Attribute
Alanine Transaminase (ALT)	Numerical	Attribute
Aspartate Aminotransferase (AST)	Numerical	Attribute
Bilirubin (BIL)	Numerical	Attribute
Choline esterase (CHE)	Numerical	Attribute
Cholesterol (CHOL)	Numerical	Attribute
Creatine (CREA)	Numerical	Attribute
Gamma-glutamyl Transferase (GGT)	Numerical	Attribute

In TABLE 1, there are several columns, ranging from age to gamma-glutamyl transferase (GGT) as the independent variable and the category column as the dependent variable. The data will be analyzed using machine learning.

B. *DATA PRE – PROCESSING*

Data pre-processing is a crucial stage. The data that has been obtained should first go through this stage. Data pre-processing converts raw data into quality data to be processed at the next stage [24]. Several stages are carried out in this process, namely, label encoding, handling missing values, and normalizing data.

1. Label Encoding

Label encoding is a method used in data processing to convert labels or categories into numerical representations to analyze and create machine-learning models [25]. In this research, the label encoding process converts categorical features such as label and gender in the dataset into numerical form [26]. Raw data is often not suitable or ready for model building.

2. Handling Missing Value

Handling missing values is an essential step in data analysis, which involves strategies to address missing values in a dataset, such as replacing missing values with reasonable estimates or deleting incomplete data [27]. In this research, the process of handling missing values is to replace the unavailable values with the average (mean) value of the relevant attributes [28].

3. Data Normalization

Data normalization is the process of rescaling values in a dataset so that they fall within a specific range, often between 0 and 1, to facilitate more effective data analysis and machine learning [29]. In this research, the data normalization process is to equalize the scale of values between variables, which also improves the model's accuracy because, with the same value, the model will recognize the data more efficiently [30]. The following is the formula for min-max scaling as shown in (1).

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

C. *FEATURE SELECTION*

Feature selection is a process in data analysis that reduces data dimensionality, improves computational efficiency, removes attributes that do not contribute significantly to the analysis or modeling task, and prevents overfitting [31]. The data used is categorical, where feature selection can be done using the chi-square method.

The Chi-Square method in feature selection is one of the techniques used to measure the relationship between category attributes (features) in a dataset and category target variables [20]. It helps identify features that have a strong relationship with the target variable and can be used to predict or explain the target variable. This method is generally used for classification problems where the target variable is a category or class variable [32]. The following is the formula of Chi-Square as shown in (2).

$$x_c^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} \tag{2}$$

The formula explains that c represents the degrees of freedom for each variable, O_i represents the observed value in cell i, and E_i represents the expected value in cell i.

D. *CLASSIFICATION C5.0 ALGORITHM*

In this research, the development of this model is classified using the C5.0 decision parallel algorithm. The C5.0 algorithm is an algorithm in data mining and machine learning used to classify tree models [33]. This algorithm is an evolution of its predecessor, proposed to improve the performance and capability of data analysis. This algorithm constructs a hierarchical multilevel tree that can be used to classify data. The process involves selecting the best attribute as a model separator in the tree based on criteria such as information gain and gain ratio. One of the main features of C5.0 is the ability to perform pruning, i.e., removing branches of the tree that are not significant or contribute negatively to the model's performance to avoid overfitting [34]. The C5.0 algorithm is widely used in various applications such as classification, prediction, and data analysis and has become one of the most essential tools in data analysis models.

The C5.0 algorithm can also resolve missing data, a crucial feature in practical data analysis where data is often incomplete [35]. In addition, this algorithm can also group variables or attributes to solve problems involving many characteristics with a high degree of complexity. The result of using the C5.0 algorithm is a tree model that can derive clues or classify new data based on attributes learned from training data [36]. With its powerful ability to solve classification and prediction problems, the C5.0 algorithm is very useful in data analysis modeling and machine learning. Several calculations are used in building this model, namely information gain and entropy. Information Gain is the value of information to measure a data set's diversity (heterogeneity) level. The following is the formula for information gain [37] as shown in (3).

$$I(S_1, S_2, \dots, S_m) = -\sum_{i=1}^{m} P_i log_2(P_i) \qquad (3)$$

where I is the information of all cases, S represents the number of instances, Si is the number of cases in class i, m is the number of classes, and Pi is the proportion of class i, Si / S. If you are wondering why log base two, it is because log base two is commonly used in the context of computer engineering, most computer languages contain natural logarithms with base two. Then, entropy is used to measure the propensity of a class from a set of data. Here is the formula for entropy as shown in (4).

$$E(A) = \sum_{i,j=1}^{my} \frac{S_{1j} + \cdots + S_{mj}}{S} X1(S_{1j}, S_{2j}, \dots, S_{mj}) \qquad (4)$$

with E(A) as the entropy of attribute A, and Sij is the number of cases in class i and category j of attribute A. Then, the two calculations are combined by summing up, making it a gain value. Here is the formula for gain as shown in (5).

$$Gain(A) = I(S_1, S_2, \dots, S_m) + E(A) \qquad (5)$$

This algorithm processes attributes using Information Gain and Entropy. Then, the attribute with the highest Gain value is selected as the root node or root node. Then, other attributes will be evaluated similarly to get the next root node. The process is carried out until the sample category can no longer be separated. This algorithm is iterative and will check all possible splitting values.

E. *EVALUATION*

Evaluation is needed to see how robust the model is. In this research, the problem solved is a classification problem, so the appropriate evaluation uses a confusion matrix. The confusion matrix is a handy tool for evaluating the performance of classification models, especially in situations where the positive and negative classes have unbalanced proportions [38]. With the information provided by this matrix, we can understand where the model tends to go wrong and decide on appropriate actions to improve the model's performance.

The confusion matrix allows us to calculate several model performance evaluation metrics, including accuracy, precision, sensitivity, recall, specificity, F1-score, and many others. Here is the formula of the confusion matrix (TABLE 2).

TABLE 2
Tabel confusion matrix

		Prediction	
Actual		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

Accuracy measures the extent to which the model is correct in classifying the data, while precision measures the extent to which the model's optimistic predictions are correct [39]. Here is the formula for accuracy as shown in (6).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (6)$$

Precision is the ratio of the predicted positive observations on the right to the total predictable positive observations [39]. Here is the formula for precision as shown in (7).

$$Precision = \frac{TP}{TP + FP} \qquad (7)$$

Recall can be called sensitivity, the ratio of correctly predicted positive observations to all observations in the actual class [39]. Here is the formula for recall as shown in (8).

$$Recall = \frac{TP}{TP + FN} \qquad (8)$$

Where TP is the number of cases accurately predicted as positive by the model, indicating the model's ability to identify positive outcomes correctly. FP refers to cases where the model incorrectly predicted an outcome as positive when it was actually negative, indicating the model's error in identifying a negative outcome as positive. FN are cases that the model predicted as negative but were actually positive, indicating the model's failure to identify positive outcomes. TN, on the other hand, is the number of cases that were correctly predicted as negative, indicating the model's ability to identify negative outcomes accurately [40].

Area Under Curve (AUC) will be used to evaluate the effectiveness of the model in the early detection of hepatitis C disease based on a dataset of hepatitis C patients. AUC was chosen as the evaluation method because it is more suitable for assessing the value of prediction performance using both balanced and unbalanced data sets [41]. Here is the formula for AUC as shown in (9).

$$AUC = \frac{1 + T.Pr - F.Pr}{2} \qquad (9)$$

The formula explains that T is the number of positive pairs correctly classified (True Positive), F is the number of negative pairs incorrectly classified as positive (False Positive), and Pr is the number of true positives.

III. RESULT

In implementing the C5.0 decision tree algorithm classification method in the early detection of Hepatitis C disease, several stages are carried out to achieve optimal accuracy in the model built.

A. *DATA PREPROCESSING*

In this research, some features must be converted into binary form, namely the sex and category features. Where the sex features whose initial values are female and male are changed to binary numbers, namely 0 and 1, while for the category or label features, Healthy patients and suspected patients are changed to 0 and 1 with (1).

Furthermore, several features in the dataset need values added. So, to overcome this, the missing data on the feature is filled in with the average value of the feature itself. The average value of each feature that has missing values is as follows. From the TABLE 3 presented, it can be concluded that there are several features with missing values, namely ALP (18 missing values) and CHOL (10 missing values), and apart from that, there is only one missing value for each other feature. The average (mean) of each feature is as follows: ALB has an average of 41.6, ALP has an average of 68.28, ALT has an average of 28.45, CHOL has an average of 5.36, and PROT has an average of 72.04.

TABLE 3
Handling missing value

Feature	Missing Value	Mean
ALB	1	41.6
ALP	18	68.28
ALT	1	28.45
CHOL	10	5.36
PROT	1	72.04

Data normalization is performed to facilitate comparisons between variables with different value ranges, minimize outliers' impact, and improve machine learning algorithms' performance.

B. *FEATURE SELECTION*

Feature selection is done to determine the features that are most relevant to the label to improve the model's efficiency and effectiveness. This research uses chi-square to determine the best features in the dataset. The following are the results of the chi-square calculation of the features in the dataset against the Category label with (2).

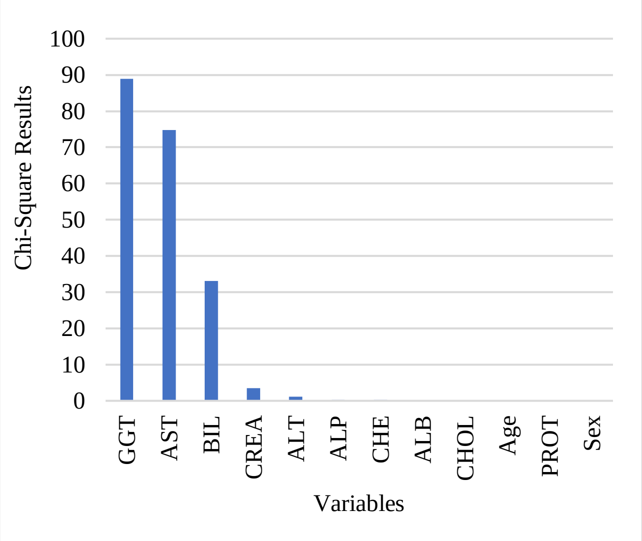

FIGURE 2. Result Selection Feature Chi-Square

FIGURE 2 shows the feature that has the highest chi-square value. This is because, in feature selection, the selected feature is the one that is most dependent on the label. In the chi-square calculation, if the two features are independent, the expected frequency is very close to the reality frequency so that the chi-square value will be low. So, the higher the chi-square value, the more dependent the feature is on the label and can be selected to be the feature used in the modeling stage. So, five features were chosen with the highest chi-square value or can be said to be the most relevant: AST, ALT, BIL, CREA, and GGT.

C. *CLASSIFICATION WITH C5.0 ALGORITHM*

With the Chi-square method, the data to be processed will be randomly weighted by determining the minimum and maximum weight values. Each element will have its weight in the dataset, and the C5.0 algorithm will be applied, and the accuracy rate will be calculated. After all particles have been calculated, the element with the best accuracy value will be found. The other particles will randomly move toward the best element in each subsequent iteration to find a better weight. This process continues until it reaches the specified iteration limit. The following are the decision tree results from the Chi-square and C5.0 models.

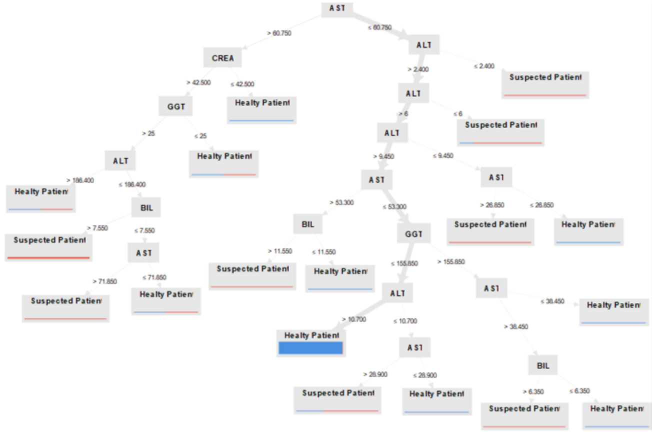

FIGURE 3. Models C5.0 Algorithm and Chi-Square

FIGURE 3 shows two colors: red, the minimum value, and blue, the maximum value of a dataset distribution. As we already know, the c5.0 algorithm produces rules that can be used to make decisions from the input data. The following is an explanation of the existing regulations.

R1 : if AST > 60.750 and CREA <= 42.500 then class = Health Patient

R2 : if AST > 60.750 and GGT < 25 then class = Health Patient

R3 : if AST > 60.750 and CREA > 42.500 and GGT > 25 and ALT > 186.400 then class = Health Patient

R4 : if AST > 60.750 and CREA > 42.500 and GGT > 25 and ALT <= 186.400 and BIL > 7.550 then class Suspected Patient

R5 : if AST > 60.750 and CREA > 42.500 and GGT > 25 and ALT <= 186.400 and BIL <= 7.550 and AST > 71.850 then class Suspected Patient

R6 : if AST > 60.750 and CREA > 42.500 and GGT > 25 and ALT <= 186.400 and BIL <= 7.550 and AST <= 71.850 then class Health Patient

R7 : if AST <= 60.750 and ALT <= 2.400 then class = Suspected Patient

R8 : if AST <= 60.750 and ALT > 2.400 and ALT <= 6 then class = Suspected Patient

R9 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT <= 9.450 and AST > 26.850 then class = Suspected Patient

R10 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT <= 9.450 and AST <= 26.850 then class = Health Patient

R11 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST > 53.300 and BIL > 11.550 = Suspected Patient
R12 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST > 53.300 and BIL <= 11.550 = Health Patient
R13 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT <= 155.850 and ALT > 10.700 then class = Health Patient
R14 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT <= 155.850 and ALT <= 10.700 and AST > 28.900 then class = Suspected Patient
R15 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT <= 155.850 and ALT <= 10.700 and AST <= 28.900 then class = Health Patient
R16 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT > 155.850 and AST <= 38.450 then class = Health Patient
R17 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT > 155.850 and AST > 38.450 and BIL > 6.350 then class Suspected Patient
R18 : if AST <= 60.750 and ALT > 2.400 and ALT > 6 and ALT > 9.450 dan AST <= 53.300 and GGT > 155.850 and AST > 38.450 and BIL <= 6.350 then class Health Patient

D. *EVALUATION*

The confusion matrix model will form a matrix consisting of true positives or positive tuples and true negatives or negative tuples, then input the prepared testing data into the confusion matrix so that the results are obtained in the table below.

TABLE 4
Confusion matrix

	True Health Patient	True Suspected Patient	Class Precision
Health Patient	108	4	96.43%
Suspected Patient	0	11	100%
Class Recall	100%	73.33%	
Accuracy		96.75%	

TABLE 4 of the testing data there are details the number of True Positive (TP) 108, False Negative (FN) 4, False Positive (FP) 0, and True Negative (TN) 11 with an accuracy of 96.75%, precision of 96.43% and recall 100%. Furthermore, the results of testing the testing data for Chi square-algorithm C5.0 against the AUC value are known in the figure below. FIGURE 4 shows the ROC graph with an AUC (Area Under Curve) value of 87.2% and a Fair Classification diagnostic accuracy level for the C5.0 Algorithm model with chi-square feature selection.

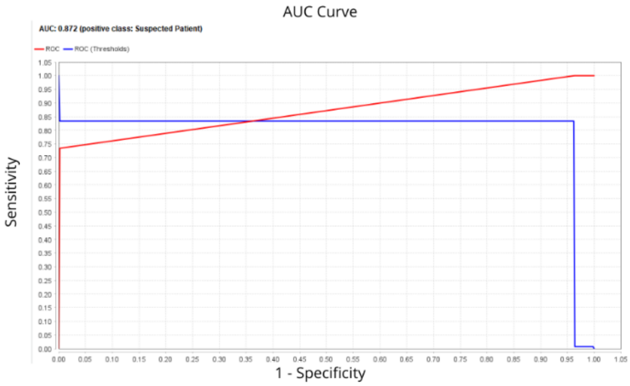

FIGURE 4. **Value Area Under Curve (AUC)**

IV. DISCUSSION

The integration of the C5.0 algorithm with Chi-Square feature selection for hepatitis C patient data has been successfully implemented in this study. This innovative approach has streamlined the diagnostic process and significantly enhanced its accuracy. By meticulously analyzing and processing the dataset from Kaggle, this methodology has set a new standard for using machine learning in medical diagnostics, particularly in the early detection of hepatitis C, a critical step in preventing the progression of the disease.

Among the myriad of features available in the dataset, five were pinpointed as the most influential through the Chi-Square feature selection process: AST (Aspartate Aminotransferase), ALT (Alanine Aminotransferase), BIL (Bilirubin), CREA (Creatinine), and GGT (Gamma-Glutamyl Transferase). These features are pivotal, for they are directly related to liver function and health, making them indispensable markers for hepatitis C. AST and ALT are enzymes found in liver cells that leak into the bloodstream during liver damage. BIL is a substance produced by the liver and indicates liver function. CREA measures kidney function but can indicate overall health, including liver health. GGT is another enzyme that, when elevated, suggests liver disease [42]. Their selection underscores the model's ability to focus on clinically relevant variables, enhancing its applicability and reliability. In implementing the model, we can see a comparison study before. The findings of this study, as shown in TABLE 5, showcase a notable accuracy rate of 96.75%, significantly surpassing those of previous research efforts in this domain. This superior accuracy is a testament to the efficacy of combining the C5.0 algorithm with Chi-Square feature selection in enhancing the predictive capability of machine learning models for hepatitis C detection. Such advancements contribute to the academic field by providing a robust model for disease prediction and offering practical implications for healthcare professionals. By adopting this model, they can achieve a more accurate diagnosis early, thereby improving patient outcomes through timely and targeted interventions. This research, therefore, not only marks a significant step forward in the application of machine learning in healthcare but also sets the stage for future innovations in the diagnosis and treatment of hepatitis C.

TABLE 5
Related research

Study	Algorithm	Result
Nivaan and Emanuel [43]	Regression Logic	Accuracy 83.33%
Hashem et al. [44]	Decision Tree Learning Algorithm	Accuracy 84.8%
Yulhendri et al. [45]	Naïve Bayes	Accuracy 86.04%
Ling Ma et al. [46]	XGBoost algorithm	Accuracy 91.56%
Farooq [47]	Random Forest Algorithm	Accuracy 93%
Purpose Methode	C5.0 + Chi-Square	**Accuracy 96.75%**

While this study offers an innovative approach with impressive accuracy in hepatitis C detection through the integration of the C5.0 algorithm with Chi-Square feature selection, there are limitations, such as the reliance on the quality of the dataset from Kaggle, which may not include broad patient demographics or incomplete data, as well as the potential neglect of interactions between variables that could provide additional insights. However, the implications are significant, offering a more efficient and accurate diagnostic method for early detection of hepatitis C, which could improve patient outcomes through timely and targeted interventions and drive further innovation in the application of machine learning in medical diagnostics and healthcare.

V. CONCLUSION

The successful application of the C5.0 algorithm combined with Chi-Square feature selection on hepatitis C patient data has accomplished the research objectives, demonstrating the feasibility and effectiveness of this approach in medical diagnostics. By carefully selecting the most relevant features for hepatitis C detection, this method has proven to streamline the diagnostic process and significantly enhance its accuracy. A 96.75% accuracy rate, a notable improvement over previous studies, underscores the potential of tailored machine-learning solutions in addressing complex health challenges.

This study's achievements set a solid foundation for future advancements in the field of medical diagnostics. One promising direction is the exploration of AutoML technologies, which promise to automate further and refine the algorithm selection process. AutoML's capability to periodically identify the most efficient algorithms based on evolving datasets could lead to continuous improvements in diagnostic accuracy and efficiency. This potential for ongoing optimization makes AutoML an exciting prospect for enhancing machine learning models dedicated to healthcare.

Looking beyond the current advancements, there is a hopeful anticipation for developing a web-based application that healthcare institutions could utilize. Such an application would not only facilitate the widespread application of the research findings but also ensure that the benefits of this advanced diagnostic tool are accessible to a broader audience. Implementing this technology in a user-friendly platform would empower healthcare professionals with a powerful tool for early hepatitis C detection, contributing significantly to improved patient care and outcomes. The vision of integrating cutting-edge machine learning models into practical healthcare applications represents a transformative step forward, promising a future where technology and medicine converge to offer unprecedented solutions to longstanding health challenges.

REFERENCES

[1] Alizargar, A., Chang, Y., and Tan, T., "Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques," *MDPI Journals*, vol. 10, no. 481, Apr. 2023, doi: bioengineering10040481.

[2] Andeli, N., Lorencin, I., Šegota, S. B., and Ca, Z., "The Development of Symbolic Expressions for the Detection of Hepatitis C Patients and the Disease Progression from Blood Parameters Using Genetic Programming-Symbolic Classification Algorithm," *MDPI Journals*, vol. 13, no. 574, Dec. 2022, doi: 13010574.

[3] Sedeno-Monge, V., *et al.*, "A comprehensive update of the status of hepatitis C virus (HCV) infection in Mexico—A systematic review and meta-analysis (2008–2019)," *Ann Hepatol*, vol. 20, pp. 1–11, Jan. 2021, doi: https://doi.org/10.1016/j.aohep.2020.100292.

[4] Homolak, J., *et al.*, "A Cross-Sectional Study Of Hepatitis B And Hepatitis C Knowledge Among Dental Medicine Students At The University Of Zagreb," *Acta Clin Croat*, vol. 60, no. 2, pp. 216–230, Jul. 2021, doi: 10.20471/acc.2021.60.02.07.

[5] Sachdeva, R. K., Bathla., Rani, P., Solanki, V., and Ahuja, R., "A systematic method for diagnosis of hepatitis disease using machine learning," *Innov Syst Softw Eng*, vol. 19, no. 3, pp. 71–80, Jan. 2023, doi: https://doi.org/10.1007/s11334-022-00509-8.

[6] ManeI, R., *et al.*, "Evaluation of five rapid diagnostic tests for detection of antibodies to hepatitis C virus (HCV): A step towards scale-up of HCV screening efforts in India," *Plos One Journals*, pp. 1–10, Jan. 2019.

[7] Shivkumar, M. S., Peeling, P. R., Jafari, M. Y., Joseph, P. L., and Pai, M. M. P. N. P., "Accuracy of Rapid and Point-of-Care Screening Tests for Hepatitis C," *Ann Intern Med*, vol. 157, no. 8, pp. 558–566, Oct. 2012, doi: https://doi.org/10.7326/0003-4819-157-8-201210160-00006.

[8] Leathersa, J. S., *et al.*, "Validation of a point-of-care rapid diagnostic test for hepatitis C for use in resource-limited settings," *Int Health*, vol. 11, pp. 314–315, 2019, doi: 10.1093/inthealth/ihy101.

[9] Ibrahim, I. N., *et al.*, "Towards 2030 Target for Hepatitis B and C Viruses Elimination Assessing the Validity of Predonation Rapid Diagnostic Tests versus Enzyme-linked Immunosorbent Assay in State Hospitals in Kaduna, Nigeria," *Nigerian Medical Journal*, vol. 60, no. 3, pp. 161–164, Jun. 2019, doi: 10.4103/nmj.NMJ_93_18.

[10] Mahesh, B., "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Oct. 2020.

[11] Jijo, B. T., and Abdulazeez, A. M., "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal Of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, Mar. 2021.

[12] Yağanoğlu, M., "Hepatitis C virus data analysis and prediction using machine learning," *Data Knowl Eng*, vol. 142, pp. 101–120, Nov. 2022, doi: https://doi.org/10.1016/j.datak.2022.102087.

[13] Butt, M. B., *et al.*, "Diagnosing the Stage of Hepatitis C Using Machine Learning," *J Healthc Eng*, pp. 1–8, Nov. 2021, doi: 10.1155/2021/8062410.

[14] Akella, A., and Akella, S., "Applying Machine Learning to Evaluate for Fibrosis in Chronic Hepatitis C," *medRxiv*, Nov. 2020, doi: 11.02.20224840.

[15] Rajeswaria, S., and Suthendran, K., "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud,"

Comput Electron Agric, vol. 156, pp. 530–539, 2019, doi: https://doi.org/10.1016/j.compag.2018.12.013.

[16] K.V, U., and Appavu, B. S., "C5.0 Decision Tree Model Using Tsallis Entropy and Association Function for General and Medical Dataset," *Intelligent Automation And Soft Computing*, vol. 26, no. 1, pp. 61–70, 2020, doi: DOI: 10.31209/2019.100000153.

[17] Dalal, S., *et al.*, "A precise coronary artery disease prediction using Boosted C5.0 decision tree model," *Journal of Autonomous Intelligence*, vol. 6, no. 3, pp. 1–18, Jul. 2023, doi: 10.32629/jai.v6i3.628.

[18] Ghavidel, A., Pazos, P., Suarez, R. D. A., and Atashi, A., "Predicting the Need for Cardiovascular Surgery: A Comparative Study of Machine Learning Models," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 92–106, Apr. 2024, doi: https://doi.org/10.35882/jeeemi.v6i2.359.

[19] Thakkar, A., and Lohiya, R., "Attack classifcation using feature selection techniques: a comparative study," *J Ambient Intell Humaniz Comput*, Jun. 2020, doi: https://doi.org/10.1007/s12652-020-02167-9.

[20] Turhan, N. S., "Karl Pearson's Chi-Square Tests," *Journal Academic*, vol. 15, no. 9, pp. 575–580, Sep. 2020, doi: 10.5897/ERR2019.3817.

[21] Spencer, R., Thabtah, F., Abdelhamid, N., and Thompson, M., "Exploring feature selection and classification methods for predicting heart disease," *Digit Health*, vol. 6, pp. 1–10, Dec. 2020, doi: https://doi.org/10.1177/2055207620914777.

[22] Rosidin, S., Muljono, Shidik, G. F., Fanani, A. Z., Zami, F. A., and Purwanto, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data," *International Seminar on Application for Technology of Information and Communication (iSemantic)*, Oct. 2021, doi: 10.1109/iSemantic52711.2021.9573196.

[23] Fedesoriano, "Hepatitis C Prediction Dataset," Kaggle. Accessed: Mar. 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

[24] Safdari, R., Deghatipour, A., Gholamzadeh, M., and Maghooli, K., "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intelligent Medicine*, Dec. 2021, doi: , 10.1016/j.imed.2021.12.003.

[25] Sailasya, G., and Kumari, G. L. A., "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 539–545, 2021.

[26] Hancock, J. T., and Khoshgoftaar, T. M., "Survey on categorical data for neural networks," *Journal Big Data*, vol. 7, no. 28, pp. 1–41, 2020, doi: https://doi.org/10.1186/s40537-020-00305-w.

[27] Johnson, T. F., Isaac, N. J. B., Paviolo, A., and González-Suárez, M., "Handling missing values in trait data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51–62, Aug. 2021, doi: https://doi.org/10.1111/geb.13185.

[28] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, B. M. T., and Tabona, O., "A survey on missing data in machine learning," *J Big Data*, vol. 8, no. 140, Oct. 2021, doi: https://doi.org/10.1186/s40537-021-00516-9.

[29] Singh, D., and Singh, B., "Investigating the impact of data normalization on classification performance," *Applied Soft Computing Journal*, pp. 1–23, 2019, doi: https://doi.org/10.1016/j.asoc.2019.105524.

[30] Kappal, S., "Data Normalization using Median & Median Absolute Deviation (MMAD) based Z-Score for Robust Predictions vs. Min – Max Normalization," *London Journal of Research in Science: Natural and Formal*, vol. 19, no. 4, pp. 39–44, 2019.

[31] Uma, D. K. V., Padmaja, P. J., and Vinoodhini, D., "Stacked Feature Selection and C5.0 Classification Model with Tsallis Entropy for Medical Dataset," *Journal of Pharmaceutical Negative Results* , vol. 13, no. 2, pp. 393–399, 2022.

[32] Ray, S., Alshouiliy, K., Roy, A., AlGhamdi, A., and Agrawal, D. P., "Chi-Squared Based Feature Selection for Stroke Prediction using AzureML ," *Intermountain Engineering, Technology and Computing (IETC)*, Dec. 2020, doi: 10.1109/IETC47856.2020.9249117.

[33] Tian, J., and Zhang, J., "Breast cancer diagnosis using feature extraction and boosted C5.0 decision tree algorithm with penalty factor," *Mathematical Biosciences and Engineering*, vol. 19, no. 3, pp. 2193–2205, Jan. 2022.

[34] Ayinla, I. B., and Akinola, S. O., "An Improved Collaborative Pruning Using Ant Colony Optimization and Pessimistic Technique of C5.0 Decision Tree Algorithm," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 12, pp. 111–123, Dec. 2020, doi: https://doi.org/10.5281/zenodo.4427699.

[35] Widyananda, W., Purnomo, M. F. E., Aswin, M., Mudjirahardjo, P., and Pramono, S. H., "Dataset Missing Value Handling And Classification Using Decision Tree C5.0 And K-Nn Imputation: Study Case Car Evaluation Dataset," *J Theor Appl Inf Technol*, vol. 100, no. 12, pp. 4503–4512, Jun. 2022.

[36] Pathan, P. S. S., "An Approach to Decision Tree Induction for Classification," *Turkish Journal of Computer and Mathematics Education* , vol. 12, no. 12, pp. 919–928, May 2021.

[37] Badr, S. M., "Adaptive Layered Approach using C5.0 Decision Tree for Intrusion Detection Systems (ALIDS)," *Int J Comput Appl*, vol. 66, no. 22, pp. 18–22, Mar. 2013.

[38] Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., and Debauche, O., "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Comput Sci*, vol. 191, pp. 487–492, Aug. 2021, doi: 10.1016/j.procs.2021.07.062.

[39] Krstinić, D., Braović, M., Šerić, L., and Božić-Štulić, D., "Multi-Label Classifier Performance Evaluation With Confusion Matrix," *Computer Science & Information Technology (CS & IT)*, pp. 1–14, 2020, doi: 10.5121/csit.2020.100801.

[40] Arif, N. H., Faisal, M. R., Farmadi, A., Nugrahadi, D. T., Abadi, F., and Ahmad, U. A., "An Approach to ECG-based Gender Recognition Using Random Forest Algorithm," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 107–115, Apr. 2024, doi: https://doi.org/10.35882/jeeemi.v6i2.363.

[41] Rodriguez, D., Herraiz, I., Harrison, R., Dolado, J., and Riquelme, J. C., "Preliminary Comparison of Techniques for Dealing with Imbalance in Software Defect Prediction," *Proc. 18th Int. Conf. Eval. Assess. Softw. Eng. - EASE '14*, vol. 43, pp. 1–10, May 2014, doi: https://doi.org/10.1145/2601248.2601294.

[42] Sookoian, S., and Pirola, C. J., "Liver enzymes, metabolomics and genome-wide association studies: From systems biology to the personalized medicine," *World J Gastroenterol* , vol. 21, no. 3, pp. 711–725, Jan. 2015, doi: 10.3748/wjg.v21.i3.711.

[43] Nivaan, G. V., and Emanuel, A. W. R., "Analytic Predictive of Hepatitis using The Regression Logic Algorithm," *IEEE*, pp. 106–110, Jan. 2021, doi: 10.1109/ISRITI51436.2020.9315365.

[44] Hashem, S., *et al.*, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *IEEE/ACM Trans Comput Biol Bioinform*, 2020.

[45] Yulhendri, Malabay, and Kartini, "Correlated Naïve Bayes Algorithm To Determine Healing Rate Of Hepatitis C Patients," *International Journal of Science, Technology & Management*, vol. 4, no. 2, pp. 401–410, Mar. 2023.

[46] Ling Ma, YongSheng Yang, Xin Ge, YiDan Wan, and Xin Sang, "Prediction of disease progression of chronic hepatitis C based on XGBoost algorithm," *International Conference on Robots & Intelligent System (ICRIS)*, Nov. 2020, doi: 10.1109/ICRIS52159.2020.00151.

[47] Farooq, S. A., "The Multi-Class Detection of Five Stages of Hepatitis C using the Machine Learning based Random Forest Algorithm," *2023 World Conference on Communication & Computing (WCONF)*, Jul. 2023, doi: 10.1109/WCONF58270.2023.10235157.

AUTHORS BIOGRAPHY

Mahmud originated in Barito Utara, Central Kalimantan. Since 2020, she has pursued her academic endeavors as a student of the Computer Science Department at Universitas Lambung Mangkurat. Her current area of research lies within the realm of data science. The study program offers the opportunity to cultivate her interest in data science. He has selected this particular interest due to my affinity towards data science and his profound fascination with this field. Additionally, his final project entailed conducting research that centered around the classification of the hepatitis C virus machine-learning method. The purpose of this research endeavor is to improve the accuracy of hepatitis C virus classification by integrating the C5.0 algorithm with Chi-Square feature selection, overcoming the limitations of current diagnostic approaches and potentially reducing diagnostic errors.

Irwan Budiman successfully finished his bachelor's degree in the informatics department at the Islamic University of Indonesia. Subsequently, he assumed the role of a lecturer in Computer Science at Universitas Lambung Mangkurt starting in 2008. Additionally, in 2010, he pursued a master’s degree in information systems at Diponegoro University. Currently, Irwan Budiman is the chair of the computer science study program at Universitas Lambung Mangkurat. His area of research expertise lies in Data Science.

Fatma Indriani is a lecturer in the Department of Computer Science at Lambung Mangkurat University. Her research interest is focused on Data Science. Before becoming a lecturer, she completed her undergraduate program in the Informatics Department at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia, in 2012. Her latest education is a doctorate in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The research fields she focuses on are Data Science and Bioinformatics.

Dwi Kartini received her bachelor’s and master’s degrees in computer science from the Faculty of Computer Science, Putra Indonesia "YPTK" Padang, Indonesia. She is also a lecturer in the Department of Computer Science. She instructs in various subjects such as linear algebra, discrete mathematics, research methods, and others. Her research interests include the applications of Artificial Intelligence and Data Mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia.

Mohammad Reza Faisal was born in Banjarmasin. Following his graduation from high school, he pursued his undergraduate studies in the Informatics department at Pasundan University in 1995 and later majored in Physics at Bandung Institute of Technology in 1997. After completing his bachelor's program, he gained experience as a training trainer in the field of information technology and software development. Since 2008, he has been a lecturer in computer science at Universitas Lambung Mangkurat while also pursuing his master's program in Informatics at Bandung Institute of Technology in 2010. In 2015, he furthered his education by pursuing a doctoral degree in Bioinformatics at Kanazawa University, Japan. To this day, he continues his work as a lecturer in Computer Science at Universitas Lambung Mangakurat. His research interests encompass Data Science, Software Engineering, and Bioinformatics.

Hasri Akbar Awal Rozaq is an informatics graduate student from Amikom University Purwokerto. Currently, he is continuing his postgraduate education in computer science at the Graduate School of Informatics, Gazi University, Ankara, and is an AI coach at Orbit Future Academy. His research focuses on AI, IoT, natural sensors, and signal processing. His primary research is on utilizing the bioelectric potential of plants to serve as natural sensors.

Oktay YILDIZ received his MSc degree in the Institute of Science from Gazi University in 2004 and his Ph.D. degree in the Institute of Information Sciences from Gazi University in 2012. He has been with the Computer Engineering Department at Gazi University, Ankara, Turkey, since 2009. His research interests include machine learning and data mining.

Wahyu Caesarendra, PhD. was born in Jakarta, Indonesia on January 31st, 1982. He received Bachelor of Engineering degree from Diponegoro University, Indonesia in 2005. He worked in automotive and electrical company prior to join Diponegoro University as a Lecturer in 2007. He received New University for Regional Innovation (NURI) and Brain Korea 21 (BK21) scholarships for Master study in 2008 and obtained his Master of Engineering (M.Eng) degree from Pukyong National University, South Korea in 2010. In 2011, Wahyu Caesarendra was awarded of University Postgraduate Award (UPA) and International Postgraduate Tuition Award (IPTA) from University of Wollongong. He received Doctor of Philosophy (Ph.D) degree from University of Wollongong in 2015. Now, he is Assistant Professor (Faculty of Integrated Technologies) at Universiti Brunei Darussalam, Brunei Darussalam.