Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

Implementation of Extra Trees Classifier and Chi-Square Feature Selection for Early Detection of Liver Disease

Muhammad Akmal Al Ghifari[1], Irwan Budiman[*2], Triando Hamonangan Saragih[3], Muhammad Itqan Mazdadi[4], Rudy Herteno[5], Hasri Akbar Awal Rozaq[6]

[1,2,3,4,5]Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia
[2]Graduate School of Informatics, Department of Computer Science, Gazi University, Türkiye

Email: [1]irwan.budiman@ulm.ac.id

Received : Dec 27, 2024; Revised : Feb 10, 2025; Accepted : Feb 19, 2025; Published : Oct 23, 2025

Abstract

The imbalanced distribution of medical data poses challenges in accurately detecting liver disease, which is crucial as symptoms often remain unnoticed until advanced stages. This study examines the application of the Extra Trees Classifier algorithm and chi-square feature selection for early detection of liver disease. Compared to traditional methods like Random Forest and SVM, the Extra Trees Classifier offers enhanced computational efficiency and better handling of imbalanced datasets, while chi-square feature selection helps identify the most relevant medical indicators. The data consists of five medical variables likely to be laboratory test results from patient samples, with labels indicating classes A and B. The data is randomly divided with a ratio of 80% for each class. To address data imbalance, SMOTE technique was applied before the data was randomly split into a ratio of 80% for training and 20% for testing to ensure effective learning and testing of the model's performance. The results showed that with the help of chi-square feature selection, the Extra Trees Classifier algorithm could provide fairly accurate predictions in liver disease classification, with an accuracy of 82.6%, sensitivity of 85.5%, precision of 78.3%, and F1-Score of 81.7%. These results demonstrate significant improvement over existing methods, and the proposed approach can aid healthcare practitioners in making timely diagnostic decisions, potentially reducing mortality rates through early intervention in liver disease cases.

Keywords : *Chi-Square, Extra Trees Classifier, Feature Selection, Liver.*

1. INTRODUCTION

The liver is a vital organ for humans. This organ is located in the right side of the abdominal cavity, just below the diaphragm [1]. The liver has several functions, including detoxification and neutralization of toxins, regulation of hormone circulation, and regulation of blood composition containing fats, sugars, proteins, and other substances [2]. The liver also produces bile, a substance that aids in the digestion of fats [3]. Liver disease is a disorder affecting all liver functions [4]. Liver disease is often referred to as a silent killer, because it may not cause any symptoms [5]. Liver disease is an inflammatory disease of the liver. Generally, liver disease can be caused by an unhealthy lifestyle, but other factors include congenital liver abnormalities present at birth, metabolic disorders, viral or bacterial infections, malnutrition, alcohol and other substance dependence, and smoking [6].

One of the problems facing society today is the delay in medical treatment for liver patients, as most patients only seek medical attention after the disease has reached an advanced stage [7]. Delayed diagnosis of liver disease can lead to various serious complications such as cirrhosis, liver cancer, and even death. Research shows that the mortality rate from liver disease increases by up to 50% in cases that are diagnosed late [8]. To address the problem of worsening health conditions in patients, regular checkups and prevention of the risk of chronic disease attacks are necessary. However, regular check-

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

ups and risk prevention for liver disease are not carried out by some people for several reasons, including busy schedules, the high cost of check-ups, and fear of being diagnosed with a chronic disease [9]. Regular checkups and early prevention of liver disease symptoms are essential so that patients can receive appropriate treatment. Early diagnosis of liver disease can increase patient life expectancy. The life expectancy of patients with liver disease can be improved if early diagnosis is made. To diagnose liver disease, medical professionals need to perform many tests and examinations to confirm the diagnosis, but they cannot guarantee the accuracy of the diagnosis. One of the tests and examinations used to diagnose liver disease is a liver function test [2]. Liver function tests are very helpful in diagnosing liver disease. The parameters measured in liver function tests include albumin, alkaline phosphatase, total protein, aspartate aminotransferase, alanine aminotransferase, direct bilirubin, total bilirubin, gamma-glutamyl transferase, prothrombin time, platelet count, triglycerides, and others.

Based on the need to find a more accurate method for detecting and diagnosing liver disease, researchers examined the application of a machine learning algorithm, namely the Extra Trees Classifier algorithm, and the use of chi-square feature selection for the implementation of early liver disease detection, with the aim of obtaining more accurate diagnostic information and improving the analysis of medical experts in diagnosing liver disease at an earlier stage [10]–[12], with the aim of obtaining more accurate diagnostic information and improving the analysis of medical experts in diagnosing liver disease at an earlier stage [13]. Extra Trees Classifier, also known as Extremely Randomized Trees, is a type of ensemble machine learning technique that combines the results of several uncorrelated decision trees collected in a decision tree to produce its classification results [14]. Meanwhile, chi-square feature selection is a test of independence between two qualitative variables to determine whether there is a correlation between two categorical variables [15]. In other words, this test is used to determine whether the values of one categorical variable are interdependent on the values of another categorical variable.

The application of the Extra Trees Classifier algorithm and the use of chi-square feature selection in healthcare are very important because the application of these two methods can predict patterns across data sets, enabling faster and more accurate determination of risk or diagnostic factors for diseases [16]. This method can enable early detection and prevent many cases of liver disease from developing to the point where a biopsy or complex treatment is required [17]. If liver disease is diagnosed at an early stage, it can be treated. The healthcare sector widely uses machine learning techniques [18], especially for the diagnosis and classification of certain diseases based on characteristic information. Medical experts will find it easier to make decisions about patients using this method. When data is obtained using machine learning feature engineering techniques, the raw input feature space is usually full of irrelevant feature information and often exhibits high dimensions.

Several previous studies have demonstrated the effectiveness of machine learning in diagnosing liver disease. According to Panwar et al. (2022), classification algorithms are often used to predict liver disease because they can predict whether a patient has the disease or not based on certain features or characteristics [19]. Another study conducted by Shaker Abdalrada et al., (2022), found that the use of machine learning capabilities to create useful prediction models would be very helpful in disease recognition and effective real-time medical decision-making [20].

Although various machine learning methods have been used for liver disease detection, such as Random Forest, SVM, and XGBoost, these methods have several limitations. Random Forest tends to overfit on unbalanced datasets, SVM requires long computation times for large datasets, and XGBoost is sensitive to noise and outliers [21], [22]. The Extra Trees Classifier overcomes these limitations with several advantages: (1) it is more resistant to overfitting due to a more extreme randomization process, (2) it has faster computation time because it does not need to find the optimal split point, and (3) it performs better on unbalanced datasets [23], [24]. Additionally, the use of chi-square feature selection

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

provides advantages in identifying the most relevant medical features, reducing model complexity, and improving the interpretability of diagnosis results [25].

Based on the description of the problem, this study will use the Extra Trees Classifier algorithm and chi-square in feature selection to identify liver disease classification. The combination of these two methods has rarely been explored in the context of liver disease detection, even though it has the potential to improve diagnostic accuracy while maintaining computational efficiency. This study will produce high accuracy with the aim of early detection and reducing the possibility of errors in liver disease recognition. In addition, this study will be compared with the results of previous studies to provide better context regarding the effectiveness of the methods used.

2. METHOD

This study uses one of the machine learning techniques, specifically supervised learning, namely the classification method with the Extra Trees Classifier algorithm. In this study, several stages can be seen in the following research flow.

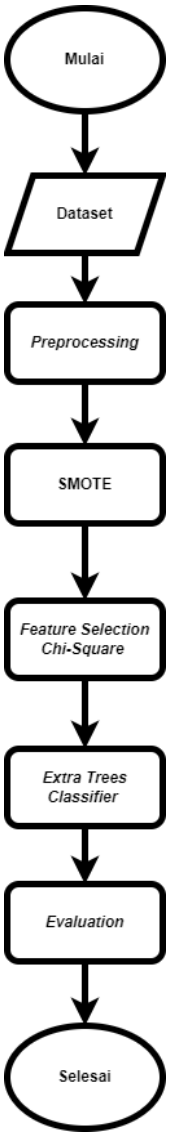

Figure 1. Research Flow

In Figure 1 above, the data collected in a dataset will be entered into the preprocessing stage, which includes handling missing values, data normalization, and label coding. After the preprocessing

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

stage is complete, SMOTE is used to balance the amount of minority class data with majority class data. The next step is feature selection using the chi-square test to determine the most relevant features. Next, the model is built using the Extra Trees Classifier algorithm, and the final step is to evaluate the model's performance using relevant metrics to ensure accuracy and effectiveness in detecting liver disease.

2.1. Dataset

The dataset used in this study is the Indian Liver Patient Dataset, obtained from the UCI Machine Learning Repository and accessible via https://www.kaggle.com/datasets/uciml/indian-*liver*-patient-records. This dataset contains patient medical records consisting of 10 features and 583 rows of data, with two class labels: 416 rows for patients with liver disease (class 1) and 167 rows for patients without liver disease (class 2). The dataset is divided into two parts, namely test data and training data, with a ratio of 80% and 20%. This ratio has proven to be effective for analysis using machine learning [26]. Detailed data for each data attribute can be seen in the following table.

Table 1. Dataset Information

Features	Data Type	Description
Age	Numerical	Attribute
Gender	Binary (0,1)	Attribute
Total Bilirubin	Numerical	Attribute
Direct Bilirubin	Numerical	Attribute
Alkaline_Phospotase	Numerical	Attribute
Alamine_Aminotransferase	Numerical	Attribute
Aspartate_Aminotransferase	Numerical	Attribute
Total_Protein	Numerical	Attribute
Albumin	Numerical	Attribute
Albumin_And_Globulin Ratio	Numerical	Attribute
Category	Numerical	Class

Table 1 shows a detailed description of the features contained in the research dataset. The table includes several columns. Age is a numerical feature that indicates the patient's age, while Gender is a binary feature with a value of 0 or 1 that indicates the patient's gender. The Total Bilirubin feature measures the total bilirubin level in the blood and is a numerical feature, as is Direct Bilirubin, which measures the direct bilirubin level in the blood. The Alkaline Phosphatase feature measures the level of alkaline phosphatase enzyme in the blood, and Alamine Aminotransferase measures the level of alamine aminotransferase enzyme in the blood. Aspartate Aminotransferase is a numerical feature that measures the level of aspartate aminotransferase enzyme in the blood, while Total Protein measures the total protein level in the blood. Albumin is a numerical feature that measures the level of albumin in the blood, and Albumin and Globulin Ratio measures the ratio of albumin and globulin in the blood. Finally, Category is a numerical feature that serves as a target class or dependent variable, with two categories indicating whether or not the patient has liver disease.

2.2. Data Preprocessing

Data pre-processing is an important step that must be done first. It converts raw data into quality data that can be processed in the next stage. Several steps carried out in this process are handling missing values, data normalization, and label encoding [27]. The first stage of missing values is one of several data quality challenges that often occur in real-world data sets [28]. This common problem usually

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

affects data analytics performance, causing high bias and low accuracy. In this study, the process of handling missing values is to replace unavailable values with the average value of the relevant attributes.

The second stage of data normalization is only necessary when the data sets have different ranges [29]. The purpose of data normalization is to convert the values of numeric columns in a data set to a common scale without changing the overall range of values. The data normalization process involves equalizing the value scales between variables, which also improves accuracy because with the same values, the model will recognize the data more efficiently. The following is the basic formula for min-max scaling as shown in the formula.

$$X_{norm} = \frac{x' - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Where x is the attribute value, min(x) and max(x) are the minimum and maximum absolute values of x, x' is the old value of each entry in the data. The third stage of label encoding is a method used in data processing to convert labels or categories into numerical representations for analysis and machine learning modeling. The purpose of label encoding is to facilitate the data mining process because data mining methods are generally better at reading numerical data.

2.3. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used in machine learning to address class imbalance in datasets [30]. Class imbalance occurs when one class (minority) has far fewer samples than the other class (majority). This can cause machine learning models to be biased towards the majority class and less effective in predicting the minority class. Class imbalance can cause significant problems in machine learning, such as model bias, where the model tends to predict the majority class more often than the minority class because the majority class is more represented in the training data, and lack of generalization, where the model cannot recognize enough patterns in the minority class, which can lead to poor predictions when faced with new data containing the minority class [31]. SMOTE works by generating synthetic samples from the minority class to increase the sample size [32].

SMOTE increases the number of data rows by generating random synthetic data for the minority class from its nearest neighbors. Since they are created based on the original characteristics of the dataset, the new data rows are more similar to the original data [33]. Here is the formula for calculating SMOTE using geometric distance.

$$D(x,y) = \sqrt{(X_1 - Y_1)^2 + .. + (X_n - Y_n)^2} \quad (2)$$

In this equation, D(x,y) is the Euclidean distance between two points, x and y. Point x represents the coordinates of the first point in n-dimensional space, with n components that can be written as (X1, X2,…,Xn). Meanwhile, point y represents the coordinates of the second point in n-dimensional space, also with n components written as (Y1, Y2, …, Yn). The component Xi is the i-th component of point x, where i ranges from 1 to n, and Yi is the i-th component of point y, where i also ranges from 1 to n. After calculating the instance distance with geometric distance, formula (3) is used to create replication data from the nearest instance.

$$Xsyn = X_i + (X_{knn} - X_i) \times \sigma \quad (3)$$

In this formula, Xsyn is the generated synthetic data point, Xi is the original data point from the selected minority class, and Xknn is the closest neighbor data point of Xi in the minority class. The variable σ is a random value between 0 and 1. The process of generating synthetic data points is done by selecting the original data point X i from the minority class and its nearest neighbor Xknn, then

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

taking the difference between Xknn and Xi, multiplying this difference by the random value σ, and adding the result to Xi. This process is repeated for each data point in the minority class until the dataset is balanced.

2.4. Feature Selection

Feature selection is a process in data analysis that reduces the dimension of data, improves computational efficiency, eliminates attributes that do not contribute significantly to the analysis or modeling task, and prevents overfitting [30]. The data used is categorical, where feature selection can be performed using the chi-square method.

The chi-square method in feature selection is one technique used to measure the relationship between categorical attributes in a dataset and categorical target variables. 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 commonly used for classification problems where the target variable is a categorical variable or class. The following is the chi-square formula as shown in the formula.

$$X^2 = [\frac{\Sigma(f_{(o)} - f_{(e)})^2}{f_{(o)}}] \quad (4)$$

Where X^2 is chi-square value, $f_{(e)}$ is the expected frequency, and $f_{(0)}$ is the frequency obtained/observed.

2.5. Extra Tree Classifier

Extra Trees Classifier is an ensemble machine learning algorithm used for classification. It belongs to the family of algorithms [30] Random Forest, which forms a number of decision trees and uses their predictions to generate the final result. Extra Trees Classifier differs from traditional Random Forest in that it uses random thresholds for each feature rather than selecting the best split to form each decision tree.

The Extra Trees Classifier algorithm uses a random splitting procedure for numerical attributes. This procedure is governed by parameters for the number of attributes randomly selected for each node and the minimum sample size for node splitting. This method is used to create an ensemble model with trees by using the original complete learning sample several times. Predictions for regression and classification are collected through majority vote or arithmetic mean. This method aims to reduce variance by explicitly randomizing cut points and attributes. Bias is minimized by using the complete learning sample. Although the node separation procedure is complex, it improves computational efficiency. Parameters and each influence attribute selection, noise average, and variance reduction. Although they can be changed, the default settings are better for method autonomy and computation. The Extra Trees Classifier algorithm is used in various applications such as classification, prediction, and data analysis and has become one of the most important tools in data analysis models.

Predictions are made in classification by majority vote from decision trees. With its powerful ability to solve classification and prediction problems, it is very useful in data analysis modeling and machine learning. The extra-tree classifier was chosen for its explicit meaning, simplicity, and easy conversion to “if-then” rules. The extra-tree method was chosen for its randomization property for numerical inputs. This idea is very useful in problems involving a large number of numerical features. It often leads to improved accuracy.

2.6. Evaluation

The tested classification model was then evaluated to determine its performance. In this study, the problem solved was a classification problem, so the appropriate evaluation used a confusion matrix

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

[34] The confusion matrix is a useful tool for evaluating the performance of classification models, especially in situations where positive and negative classes are imbalanced. With the information provided by this matrix, we can understand where the model is making mistakes and decide on the appropriate actions to improve the model's performance [35].

The confusion matrix itself consists of four parts, namely True Positive (TP), which is the number of cases where the model correctly predicts the positive class; False Positive (FP), which is the number of cases where the model incorrectly predicts a positive class when it is actually negative; True Negative (TN), which is the number of cases where the model correctly predicts a negative class; and False Negative (FN), which is the number of cases where the model incorrectly predicts a negative class when it is actually positive. Several measurement indicators are used in the evaluation, such as accuracy (the degree of correctness of the model), sensitivity/recall (the ratio of correct positive predictions to the total correct positive data), precision (the ratio of correct positive predictions to the total predicted positive results), and F1-score (the division of precision and sensitivity). Here are the formulas [36].

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

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

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

$$F1 - score = \frac{2*(precision* recall)}{(precision+recall)} \quad (8)$$

3. RESULT

3.1. Preprocessing Result

The dataset used has four empty rows of data or missing values in the Albumin_and_Globulin_Ratio feature. These values are then filled with the most frequently occurring values in the feature column. The dataset has a wide range of values between features other than missing values. The dataset is normalized so that the range of values between features is uniform to facilitate the data mining process.

Data normalization was performed using min-max because this method has the ability to overcome the weaknesses of methods such as z-score or decimal scaling. Next, the gender feature data with a value of "Female" was changed to the number 0 and the value "Male" was changed to the number 1, as can be seen in Table 2.

Table 2. Data Normalization

Features	Nilai Awal	Hasil Label Encoding
Gender	Female	0
	Male	1

Where the characteristics of gender, whose initial values are female and male, are converted to binary numbers, namely 0 and 1. Finally, the class labels in the data set that have the numbers 1 and 2 are changed to 1 for class=1 and 0 for class=2.

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

3.2. SMOTE Process

Class 1 data is more abundant than class 0 data, so oversampling is used to balance the data distribution with SMOTE. After SMOTE is applied, the distribution of data from the features becomes balanced. SMOTE oversampling applies to all features in the dataset, not just the Total_Billirubin feature. This results in a dataset with 416 rows of class 1 (positive) and 0 (negative) data, so that the amount of data in the entire dataset is balanced, as shown in Table 3.

Table 3. SMOTE Pre- and Post-Selection Data Results Based on Selector

Selector	Data Training	Data Testing
1	416	416
2	167	416

It can be concluded that there is a possibility that data imbalance occurs in the target. Therefore, the next step is to calculate the instance distance along with the geometric distance in the target, with a total of 416 data. For classification, the SMOTE result dataset is used after class balancing is complete.

3.3. Feature Selection Result

Feature selection was performed to determine the features most relevant to the label in order to refine the efficiency and effectiveness model. The following are the chi-square calculation results for each feature.

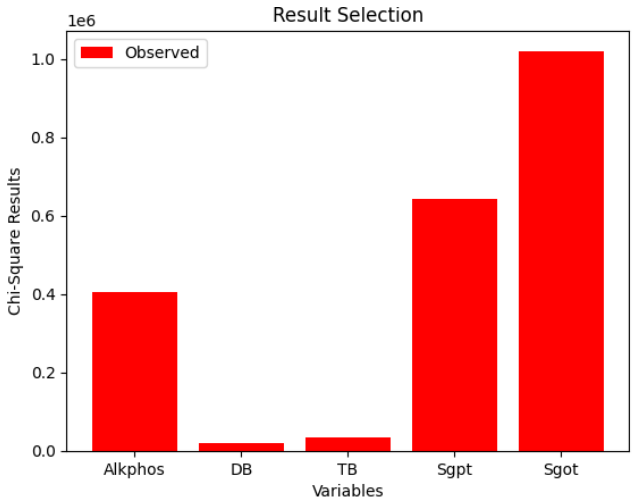

Figure 2. Chi-Square Feature Selection Result

Figure 2 shows the feature selection results using the chi-square method on a dataset that has undergone the SMOTE process to address data imbalance. The chi-square analysis results indicate that the Sgot feature has the highest value of 1.0, indicating a strong dependence on the target label (liver disease). Other features such as Sgpt with a value of 0.62, Alkphos 0.4, TB 0.12, and DB 0.11 also have significant values but are lower than Sgot. A high chi-square value indicates a large difference between the expected and actual frequencies, indicating a significant dependence between the feature and the label. Therefore, features with the highest chi-square values, such as Sgot, Sgpt, Alkphos, TB, and DB, were selected for use in modeling because they are highly relevant in predicting liver condition. This selection was based on chi-square statistical analysis, which measures the strength of the association between each feature and the target variable, ensuring that these features play a significant role in the classification of liver disease.

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

3.4. Model Performance

Like the Extra Trees Classifier, the data clustering process becomes more efficient and can produce predictive models with high accuracy and good generalization capabilities for new data. This process continues until it reaches the specified iteration limit. The following are the decision tree results from the Extra Trees Classifier and Chi-square.

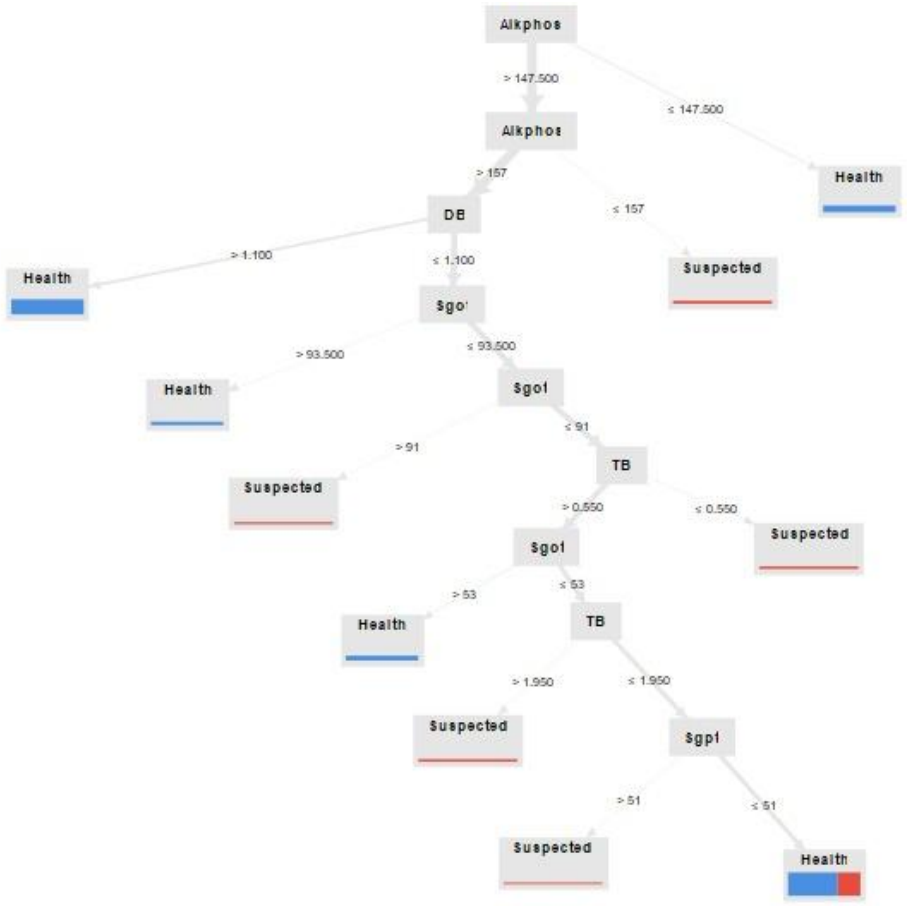

Figure 3. Extra Trees Classifier

Figure 3 shows two colors: red, minimum value, and blue, maximum value of the data set distribution. As we know, Extra Trees Classifier generates rules that can be used to make decisions from input data. The following is an explanation of the existing rules. The purpose of this analysis is to use the Extra Trees Classifier algorithm to classify medical data based on several variables. The variables used are Alkphos, DB, TB, Sgpt, and Sgot.

3.5. Evaluation

Performance measurement of the model created using a confusion matrix, where the test results can be seen in Table 4.

Table 4. Confusion Matrix Results

		Predicted	
		Positive (1)	**Negative (0)**
Actual	Positive (1)	65	18
	Negative (0)	11	73

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

The table above shows the accuracy calculated using the F1-Score value. The following are the calculations for each indicator.

$$Accuracy = \frac{(65+73)}{(65+11+73+18)} = 0{,}826 \quad (9)$$

$$Recall = \frac{65}{65+11} = 0{,}855 \quad (10)$$

$$Precision = \frac{65}{65+18} = 0{,}783 \quad (11)$$

$$F1 - score = \frac{2*(0{,}783*0{,}855)}{(0{,}783+0{,}855)} = 0{,}817 \quad (12)$$

Based on the above calculations, the accuracy obtained is 0.826, which indicates that the model is able to classify data correctly in approximately 82.6% of total cases. The sensitivity is 0.855, indicating that the model successfully identifies 85.5% of positive cases correctly. The model's precision is recorded at 0.783, meaning that of all positive predictions, 78.3% are actually positive. The F1-Score obtained is 0.817, indicating a good balance between precision and sensitivity. Overall, the Extreme Trees Classifier model shows solid and effective performance in classifying liver disease data.

4. DISCUSSIONS

The Extra Trees Classifier with Chi-Square feature selection on liver disease patient data has been successfully applied in this study. This innovative approach has simplified the diagnostic process and significantly improved its accuracy. By carefully analyzing and processing the dataset from Kaggle, this methodology has set a new standard in the use of machine learning in medical diagnostics, particularly in the early detection of liver disease, an important step in improving patient survival rates.

Among the myriad of features available in the dataset, five were selected as the most influential through the Chi-Square feature selection process: Sgot, Sgpt, Alkphos, TB, and DB. The selection of these five features proved to be more effective than previous studies that used more features. For example, Yang et al. (2023) used 10 features but only achieved an accuracy of 78.5% [4]. Reducing the number of features not only improves computational efficiency but also reduces the risk of overfitting. These characteristics are crucial because they are directly related to liver function and health, making them essential markers for liver disease. Their selection underscores the model's ability to focus on clinically relevant variables, thereby improving its applicability and reliability. In implementing the model, we can refer to previous related studies.

Table 5. Research Comparison

Previous Study	Methods	Results
[28]	SMOTE-Support Vector Machine	65%
[29]	SMOTE-XGBoost-Bayesian Search	76.7 %
[30]	SMOTE-Random Forest	77,06%
[31]	ADASYN- Support Vector Machine	80,4%
Our Model	SMOTE-Chi square-Extra Trees Classifier	82,6%

The findings of this study, as shown in Table 5, indicate the highest accuracy rate of 82.6%, surpassing previous research efforts in this domain. This superior accuracy is evidence of the effectiveness of combining the Extra Trees Classifier with Chi-Square feature selection in improving the predictive capabilities of machine learning models for early detection of liver disease. This advancement contributes to the academic field by providing a robust model for disease prediction and

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

offering practical implications for healthcare professionals. By adopting this model, they can achieve more accurate early diagnoses, thereby improving patient outcomes through timely and targeted interventions. Therefore, this research not only marks a significant step forward in the application of machine learning in healthcare but also paves the way for future innovations in the diagnosis and treatment of liver disease.

Although this study offers an innovative approach with impressive accuracy in the early detection of liver disease through the integration of the Extra Trees Classifier with Chi-Square feature selection, there are limitations, such as dependence on the quality of the data set from Kaggle, which may not be reliable. It may not cover a broad patient demographic or include incomplete data, as well as potentially ignoring 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 liver disease, which could improve patient survival rates through timely and targeted interventions and drive further innovation in the application of machine learning in medical diagnostics and healthcare.

5. CONCLUSION

The liver is a vital organ that neutralizes toxins, regulates hormone circulation, and aids in fat digestion. Liver disease, often referred to as a silent killer because it does not show early symptoms, can be caused by unhealthy lifestyles, infections, congenital disorders, alcohol addiction, and smoking. Delayed diagnosis often causes the patient's health condition to worsen, so regular checkups are necessary for early detection and proper treatment.

This study used a machine learning algorithm, Extra Trees Classifier, and chi-square feature selection to detect liver disease early. The results showed a model accuracy of 82.6%, sensitivity of 85.5%, and precision of 78.3%. This methodology improves diagnostic accuracy by selecting the most influential features, namely Sgot, Sgpt, Alkphos, TB, and DB. The combination of Extra Trees Classifier with chi-square feature selection proved to be more accurate than previous methods, such as SMOTE-Support Vector Machine.

In the future, research is expected to develop more complex models, integrate multimodal data, and improve data quality. In addition, the development of real-time diagnostic tools, longitudinal studies, and interdisciplinary collaboration are also the focus. These efforts aim to find more efficient and accurate early detection methods, improve patient life expectancy, and encourage innovation in the application of machine learning in medical diagnostics and health care.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Gazi University and Lambung Mangkurat University for their support and facilitation of this research.

REFERENCES

[1] E. Patimah, V. B. Haekal, and D. Sandya Prasvita, "Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, vol. 2, no. 1, pp. 655–659, 2021.

[2] B. N. P. Maharani, A. D. Hendriani, and P. W. P. Iswari, "Liver Cirrhosis: Pathophysiology, Diagnosis, and Management," *J. Biol. Trop.*, vol. 23, no. 1, pp. 457–463, 2023, doi: 10.29303/jbt.v23i1.5763.

[3] M. Ghosh *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intell. Autom. Soft Comput.*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/iasc.2021.017989.

[4] Z. Guo *et al.*, "A randomized-controlled trial of ischemia-free liver transplantation for end-stage

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

liver disease," *J. Hepatol.*, vol. 79, no. 2, pp. 394–402, 2023, doi: 10.1016/j.jhep.2023.04.010.
[5] M. E. Rinella *et al.*, *AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease*, vol. 77, no. 5. 2023.
[6] L. Rong, J. Zou, W. Ran, and X. Qi, "Advancements in the treatment of non-alcoholic fatty liver disease (NAFLD)," no. January, pp. 1–18, 2023, doi: 10.3389/fendo.2022.1087260.
[7] M. V. Machado, "Aerobic exercise in the management of metabolic dysfunction associated fatty liver disease," *Diabetes, Metabolic Syndrome and Obesity*, vol. 14. pp. 3627–3645, 2021, doi: 10.2147/DMSO.S304357.
[8] A. S. Afrah, "Sistem Diagnosa Penyakit Liver Menggunakan Metode Artificial Neural Network: Studi Berdasarkan Dataset Indian Liver Patient Dataset," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 308–312, Dec. 2023, doi: 10.30591/jpit.v8i3.5346.
[9] M. T. Long, M. Noureddin, and J. K. Lim, "CLINICAL PRACTICE UPDATE AGA Clinical Practice Update : Diagnosis and Management Expert Review," *Gastroenterology*, vol. 163, no. 3, pp. 764-774.e1, 2022, doi: 10.1053/j.gastro.2022.06.023.
[10] D. S. Ali and M. A. Aljabery, "Predicting Liver Cirrhosis Stages Using Extra Trees, Random Forest, and SVM with Data Mining Techniques," *Inform.*, vol. 48, no. 21, pp. 15–26, 2024, doi: 10.31449/inf.v48i21.6752.
[11] F. Muhammad *et al.*, "Liver Ailment Prediction Using Random Forest Model," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1049–1067, 2023, doi: 10.32604/cmc.2023.032698.
[12] Y. O. Daddala and K. Shaik, "Cardiovascular Disease Prediction: Employing Extra Tree Classifier-Based Feature Selection and Optimized RNN with Artificial Bee Colony," *Rev. d'Intelligence Artif.*, vol. 38, no. 2, pp. 643–653, Apr. 2024, doi: 10.18280/ria.380228.
[13] Y. Duan *et al.*, "Association of Inflammatory Cytokines With Non-Alcoholic Fatty Liver Disease," *Front. Immunol.*, vol. 13, no. May, 2022, doi: 10.3389/fimmu.2022.880298.
[14] D. Sharma, R. Kumar, and A. Jain, "Measurement : Sensors Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning," *Meas. Sensors*, vol. 24, no. September, p. 100560, 2022, doi: 10.1016/j.measen.2022.100560.
[15] M. Mahmud *et al.*, "Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 116–124, Mar. 2024, doi: 10.35882/jeeemi.v6i2.384.
[16] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 160, Jun. 2024, doi: 10.1186/s12911-024-02550-y.
[17] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach," *Chemom. Intell. Lab. Syst.*, vol. 222, no. 06, p. 104516, Mar. 2022, doi: 10.1016/j.chemolab.2022.104516.
[18] P. Theerthagiri, "Liver disease classification using histogram-based gradient boosting classification tree with feature selection algorithm," *Biomed. Signal Process. Control*, vol. 100, p. 107102, Feb. 2025, doi: 10.1016/j.bspc.2024.107102.
[19] A. Panwar, V. Bhatnagar, M. Khari, A. W. Salehi, and G. Gupta, "A Blockchain Framework to Secure Personal Health Record (PHR) in IBM Cloud-Based Data Lake," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3045107.
[20] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," *J. Diabetes Metab. Disord.*, vol. 21, no. 1, pp. 251–261, Jan. 2022, doi: 10.1007/s40200-021-00968-z.
[21] S. Qin *et al.*, "Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults," *Sci. Rep.*, vol. 13, no. 1, pp. 1–7, 2023, doi: 10.1038/s41598-023-30750-5.
[22] K. Stefanus and H. Leong, "Comparison of Random Forest Algorithm Accuracy With Xgboost Using Hyperparameters," *Proxies J. Inform.*, vol. 7, no. 1, pp. 15–23, 2024, doi: 10.24167/proxies.v7i1.12464.
[23] D. Baby, S. J. Devaraj, J. Hemanth, and M. M. Anishin Raj, "Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach," *Turkish J. Electr.*

Vol. 6, No. 5, October 2025, Page. 3925-3937
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

Eng. Comput. Sci., vol. 29, no. 8, pp. 2742–2757, 2021, doi: 10.3906/elk-2104-183.

[24] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, 2023, doi: 10.3390/biomedicines11020581.

[25] F. O. Bulucu, I. Acer, F. Latıfoğlu, and S. Içer, "Predicting liver disease using decision tree ensemble methods," *Erciyes Üniversitesi Fen Bilim. Enstitüsü Fen Bilim. Derg.*, vol. 38, no. 2, pp. 261–267, 2022.

[26] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[27] A. A. Kurniawan and M. Mustikasari, "Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 544, 2021, doi: 10.32493/informatika.v5i4.6760.

[28] R. Atiq, F. Fariha, M. Mahmud, S. S. Yeamin, K. I. Rushee, and S. Rahim, "A Comparison of Missing Value Imputation Techniques on Coupon Acceptance Prediction," *Int. J. Inf. Technol. Comput. Sci.*, vol. 14, no. 5, pp. 15–25, 2022, doi: 10.5815/ijitcs.2022.05.02.

[29] I. Huda, "Implementasi Natural Language Processing (Nlp) Untuk Aplikasi Pencarian Lokasi," *J. Nas. Teknol. Terap.*, vol. 3, no. 2, p. 15, 2021, doi: 10.22146/jntt.35036.

[30] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over - sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 2, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[31] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Syst. Appl.*, vol. 178, no. April, p. 114986, Sep. 2021, doi: 10.1016/j.eswa.2021.114986.

[32] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 305–314, 2022, doi: 10.34123/icdsos.v2021i1.240.

[33] H. M. Qasim, O. Ata, M. A. Ansari, M. N. Alomary, S. Alghamdi, and M. Almehmadi, "Hybrid Feature Selection Framework for the Parkinson Imbalanced Dataset Prediction Problem," *Medicina (B. Aires).*, vol. 57, no. 11, p. 1217, Nov. 2021, doi: 10.3390/medicina57111217.

[34] D. Salirawati, "Identifikasi Problematika Evaluasi Pendidikan Karakter di Sekolah," *J. Sains dan Edukasi Sains*, vol. 4, no. 1, pp. 17–27, 2021, doi: 10.24246/juses.v4i1p17-27.

[35] K. M. Elistiana, B. A. Kusuma, P. Subarkah, and H. A. A. Rozaq, "Improvement of Naive Bayes Algorithm in Sentiment Analysis of Shopee Application Reviews on Google Play Store," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1431–1436, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1486.

[36] W. Nengsih, "Analisa Akurasi Permodelan Supervised Dan Unsupervised Learning Menggunakan Data Mining," *Sebatik*, vol. 23, no. 2, pp. 285–291, 2019, doi: 10.46984/sebatik.v23i2.771.