Type of Contribution:
Research Paper
Review Paper
Case Study

ENERGY: JURNAL ILMIAH ILMU-ILMU TEKNIK
Special Issue 2025 pp 371-388
DOI: 10.51747/energy.si2025.253

E-ISSN: 2962-2565

This article contributes to:

The Application of Machine Learning in Liver Disease Diagnosis: Analysis of Algorithm Perfomance and Axiological Implications

Sri Farida Utami[1]*, **Syaad Patmanthara**[1]

[1] Department of Electrical Engineering and Informatics, State University of Malang, 65114, **Indonesia**

*sri.farida.2505349@students.um.ac.id

Abstract

Liver disease remains a significant global health challenge, requiring accurate and timely diagnosis to improve patient outcomes and reduce healthcare costs. This study investigates the application of four machine learning classification algorithms—Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN)—to predict the presence of liver disease using a dataset sourced from Kaggle. These algorithms were evaluated based on performance metrics such as accuracy, precision, recall, and F1 score. Both Decision Tree and Random Forest achieved the highest accuracy rate of 72.41%, demonstrating their robustness in classifying liver disease cases. However, these models showed some limitations in identifying patients without liver disease. Naïve Bayes, with an accuracy of 60.34%, exhibited an impressive recall rate of 96.97%, indicating its potential in detecting liver disease cases, though at the cost of lower precision. KNN, with an accuracy of 70.69%, proved to be a competitive option in the classification task. Beyond technical performance, the study also explores the ethical and axiological implications of using machine learning in healthcare, emphasizing the importance of fairness, transparency, and human oversight. The research highlights the need for responsible deployment of machine learning technologies, ensuring they are aligned with ethical standards to avoid biases and enhance healthcare outcomes. This study demonstrates that machine learning can significantly support liver disease diagnosis, though it must be integrated with a comprehensive ethical framework to ensure equitable and transparent decision-making in clinical practice.

Keywords: Machine Learning; Liver Disease Diagnosis; Decision Tree; Random Forest; Ethical Considerations; Axiological Perspective

Article Info
Submitted: 2025-10-25
Revised: 2025-12-21
Accepted: 2025-12-25
Published: 2025-12-30

Publisher
Universitas Panca Marga

1. Introduction

Liver disease is a complex medical condition that requires precise and accurate diagnosis for effective treatment and management [1], [2], [3]. With the growing prevalence of liver diseases globally, early detection has become a critical component in improving patient outcomes and reducing healthcare costs. Traditionally, liver disease diagnosis relies heavily on clinical expertise and laboratory tests. However, the emergence of information technology and data mining has opened new avenues for enhancing diagnostic accuracy and efficiency. Machine learning, a subset of artificial intelligence, has proven to be a powerful tool in analyzing large medical datasets to identify patterns and make predictions, making it an essential asset in the diagnosis of various health conditions, including liver disease [4].

In this study, we apply several classification algorithms—Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN)—to a dataset of liver disease patients [5], [6], [7], [8]. These algorithms are commonly used in data mining for medical diagnoses, with their ability to learn from historical patient data to predict future cases. Decision Tree and Random Forest, as tree-based models, are known for their interpretability and high performance in classification tasks. Naïve Bayes, which applies Bayes' theorem with a probabilistic approach, is known for its simplicity and efficiency, particularly when dealing with independent features. K-Nearest Neighbors, a non-parametric method, is often employed for its ability to identify patterns based on data proximity, making it suitable for medical datasets where patterns can be subtle but significant.

While previous studies have explored the application of machine learning in liver disease diagnosis, there is often a focus on the technical performance of these algorithms without fully considering their ethical and humanitarian implications [9], [10]. This paper aims to bridge that gap by not only evaluating the accuracy and efficiency of these algorithms but also exploring the axiological implications of using such technologies in the medical field. Axiology, the philosophical study of values, plays an important role in ensuring that the algorithms used in healthcare applications are designed and implemented responsibly [11]. It is essential that these technologies prioritize fairness, transparency, and minimize biases to ensure that they contribute positively to healthcare outcomes without reinforcing existing inequalities.

Thus, this study goes beyond just measuring performance metrics like accuracy, precision, and recall. We also discuss the importance of selecting appropriate algorithms to minimize diagnostic errors, ensure fairness in predictions, and align the technology with ethical standards. The integration of machine learning in medical decision-making holds great promise in accelerating the diagnostic

process and improving the quality of healthcare. However, it is equally important to consider the broader societal and ethical implications of relying on these systems, particularly in fields as critical as disease diagnosis.

2. Methods

This research follows an experimental approach aimed at evaluating the effectiveness of various machine learning algorithms in classifying liver disease cases based on a dataset of patients. The study utilizes a series of classification algorithms—Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN)—to identify patterns in the dataset and predict the likelihood of liver disease presence [4], [12]. The methodological steps in this study include data collection, data preprocessing, model training, and performance evaluation, which are essential for understanding both the technical performance of these algorithms and their practical implications for clinical decision-making, as shown in **Figure 1**.

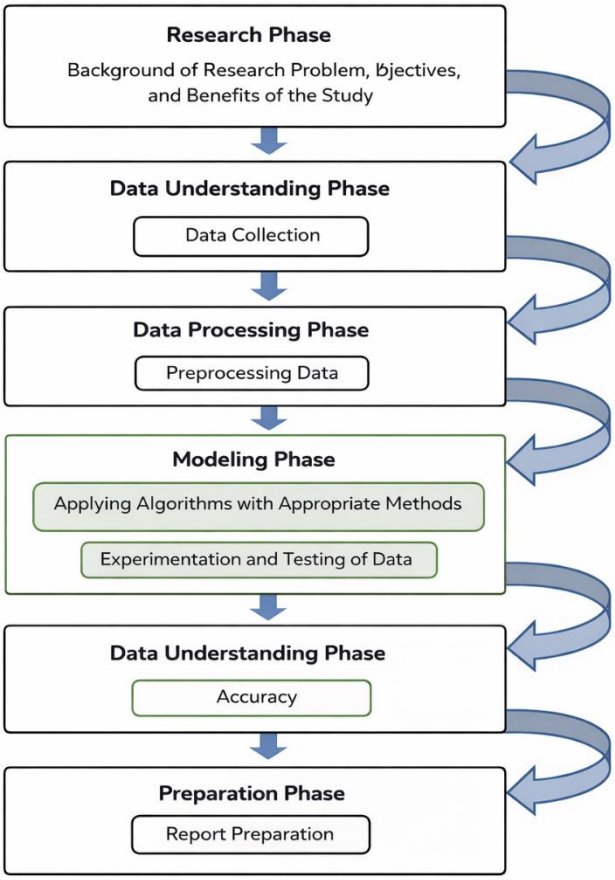

Figure 1. Research Framework

2.1 Data Collection

The dataset used in this study was sourced from Kaggle, originally compiled by Bendi Venkata Ramana and Prof. M. S. Prasad from Andhra University, India [13]. This dataset contains several attributes related to patient physiology, including demographic information (e.g., age, gender) and medical test results (e.g., bilirubin levels, liver enzymes, and protein ratios). These attributes serve as the primary input

for the machine learning algorithms, providing both the features for classification and the target labels (patients with liver disease or not). It is crucial to ensure that the dataset is both accurate and representative of the population to ensure the reliability and validity of the predictions made by the algorithms.

Table 1. Detailed 10 key attributes of the dataset.

Attribute	Description	Role in Diagnosis
Age	The age of the patient, categorized into different ranges.	Age is an important factor in diagnosing liver disease, as liver function can be influenced by age.
Gender	Binary indicator of patient sex (1 for male, 0 for female).	Gender differences can affect liver disease prevalence and prognosis, so it helps in classification.
Total_Bilirubin	The total level of bilirubin in the blood.	High bilirubin levels can indicate liver dysfunction or bile duct obstructions.
Direct_Bilirubin	The direct bilirubin level in the blood.	Elevated direct bilirubin levels are a key marker of liver disease and jaundice.
Alkaline_Phosphatase	The levels of alkaline phosphatase enzyme in the blood.	Abnormal levels are commonly associated with liver disease and can indicate liver damage.
Alamine_Aminotransferase (ALT)	A liver enzyme that helps break down proteins.	Elevated ALT levels are often used as an indicator of liver inflammation or damage.
Aspartate_Aminotransferase (AST)	An enzyme in the liver involved in amino acid metabolism.	AST levels, especially when elevated, are indicative of liver damage and are used in diagnosis.
Total_Proteins	The total amount of protein in the blood.	Protein levels reflect liver function and can help diagnose conditions like cirrhosis or liver failure.
Albumin	The protein produced by the liver that helps in maintaining blood volume.	Low albumin levels may indicate liver damage, as the liver is responsible for albumin production.
Albumin_and_Globulin_Ratio	The ratio between albumin and globulin in the blood.	This ratio is useful in diagnosing liver and kidney diseases, as an imbalance may signal liver disease.

2.2 Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for model training and ensuring that the data is in a format suitable for machine learning algorithms.

The preprocessing steps in this study include several sub-processes. First, data cleaning was performed to handle missing or inconsistent values. Missing values in attributes like liver enzyme levels or bilirubin readings were addressed using mean imputation, which is essential for ensuring that the dataset is complete for training the models. Next, data transformation was conducted to normalize continuous variables, such as bilirubin levels and albumin ratios, so they would fall within a similar range. This is particularly important for algorithms like KNN that rely on distance-based measurements. Feature selection was also carried out to identify the most relevant variables that contribute to the classification task. By selecting the most relevant features, we reduce the dimensionality of the dataset, which helps prevent overfitting and improves algorithm performance. Finally, data splitting was done using a 70-30 split, where 70% of the data was used for training and 30% for testing. Cross-validation with a 10-fold technique was also applied to ensure robust evaluation and reduce the risk of overfitting.

2.3 Algorithm Implementation

This study implemented four machine learning classification algorithms: Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN) [14]. These algorithms were chosen for their diverse approaches and strengths in classification tasks, with each having unique characteristics that influence their performance. The Decision Tree algorithm builds a tree-like model where each branch represents a decision rule based on attribute values, and each leaf represents an outcome. The model was built using the Gini index as the criterion for splitting nodes, which helps the model learn rules based on the dataset. Random Forest, an ensemble method, creates multiple decision trees and combines their predictions. This method reduces the risk of overfitting by averaging the results from several trees, which are each built on random subsets of the data. Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, which assumes feature independence. Despite this simplifying assumption, it performs well on high-dimensional datasets like medical records, and in this study, the Gaussian Naïve Bayes variant was used. Finally, the K-Nearest Neighbors (KNN) algorithm is a non-parametric method that classifies data based on the majority class of its nearest neighbors. The algorithm uses Euclidean distance as the similarity metric, and a 5-nearest neighbor approach was chosen to balance accuracy with computational efficiency.

2.4 Performance Evaluation

To assess the performance of the machine learning models, several evaluation metrics were used. Accuracy is the overall proportion of correct predictions (true positives and true negatives) made by the model compared to the total number of predictions. Precision measures the proportion of true positive predictions out of all

instances predicted as positive by the model. Recall, also known as sensitivity, measures the model's ability to correctly identify positive cases, which is crucial in medical diagnoses to avoid false negatives. The F1 score provides a harmonic mean of precision and recall, offering a balanced measure of the model's ability to classify both positive and negative cases accurately [15]. These evaluation metrics were used to compare the performance of the four algorithms in classifying liver disease cases.

3.Results and Discussion

3.1 Data Collection

The dataset utilized in this study was sourced from Kaggle, originally compiled by Bendi Venkata Ramana and Prof. M. S. Prasad from Andhra University, India [16]. This dataset is publicly available and provides an invaluable resource for analyzing liver disease diagnosis through machine learning models. It includes a variety of medical attributes that are commonly used for predicting liver disease in patients. These attributes, which include both physiological and biochemical markers, offer a comprehensive overview of the factors influencing liver health.

Additionally, the target label in the dataset is the Dataset column, where patients are classified as having liver disease or not. The dataset uses a binary classification system, where the label "1" indicates the presence of liver disease, and "2" indicates the absence.

	Age	Gender	Total_Bili	Direct_Bil	Alkaline_I	Alamine_/	Aspartate	Total_Pro	Albumin	Albumin_	Dataset
2	65	0	0.7	0.1	187	16	18	6.8	3.3	0.9	1
3	62	1	10.9	5.5	699	64	100	7.5	3.2	0.74	1
4	62	1	7.3	4.1	490	60	68	7	3.3	0.89	1
5	58	1	1	0.4	182	14	20	6.8	3.4	1	1
6	72	1	3.9	2	195	27	59	7.3	2.4	0.4	1
7	46	1	1.8	0.7	208	19	14	7.6	4.4	1.3	1
8	26	0	0.9	0.2	154	16	12	7	3.5	1	1
9	29	0	0.9	0.3	202	14	11	6.7	3.6	1.1	1
10	17	1	0.9	0.3	202	22	19	7.4	4.1	1.2	2
11	55	1	0.7	0.2	290	53	58	6.8	3.4	1	1
12	57	1	0.6	0.1	210	51	59	5.9	2.7	0.8	1
13	72	1	2.7	1.3	260	31	56	7.4	3	0.6	1
14	64	1	0.9	0.3	310	61	58	7	3.4	0.9	2
15	74	0	1.1	0.4	214	22	30	8.1	4.1	1	1
16	61	1	0.7	0.2	145	53	41	5.8	2.7	0.87	1
17	25	1	0.6	0.1	183	91	53	5.5	2.3	0.7	2
18	38	1	1.8	0.8	342	168	441	7.6	4.4	1.3	1
19	33	1	1.6	0.5	165	15	23	7.3	3.5	0.92	2
20	40	0	0.9	0.3	293	232	245	6.8	3.1	0.8	1
21	40	0	0.9	0.3	293	232	245	6.8	3.1	0.8	1
22	51	1	2.2	1	610	17	28	7.3	2.6	0.55	1
23	51	1	2.9	1.3	482	22	34	7	2.4	0.5	1
24	62	1	6.8	3	542	116	66	6.4	3.1	0.9	1
25	40	1	1.9	1	231	16	55	4.3	1.6	0.6	1
26	63	1	0.9	0.2	194	52	45	6	3.9	1.85	2
27	34	1	4.1	2	289	875	731	5	2.7	1.1	1
28	34	1	4.1	2	289	875	731	5	2.7	1.1	1
29	34	1	6.2	3	240	1680	850	7.2	4	1.2	1
30	20	1	1.1	0.5	128	20	30	3.9	1.9	0.95	2
31	84	0	0.7	0.2	188	13	21	6	3.2	1.1	2
32	57	1	4	1.9	190	45	111	5.2	1.5	0.4	1
33	52	1	0.9	0.2	156	35	44	4.9	2.9	1.4	1
34	57	1	1	0.3	187	19	23	5.2	2.9	1.2	2
35	38	0	2.6	1.2	410	59	57	5.6	3	0.8	2
36	38	0	2.6	1.2	410	59	57	5.6	3	0.8	2
37	30	1	1.3	0.4	482	102	80	6.9	3.3	0.9	1
38	17	0	0.7	0.2	145	18	36	7.2	3.9	1.18	2

Figure 1. Liver Disease Dataset

The dataset was collected with the intention of enabling machine learning algorithms to learn from these variables and predict the likelihood of liver disease. The sample size of the dataset is adequate, containing n=583 instances, which provides a solid foundation for training and testing the predictive models. This dataset is crucial for the study as it directly links the physiological and biochemical factors to the presence or absence of liver disease, making it highly relevant for evaluating the performance of classification algorithms.

However, like any real-world dataset, it has inherent limitations. There may be missing or inconsistent data, especially in clinical datasets where certain medical tests might not have been performed on every patient. Furthermore, the dataset's demographic composition could potentially introduce biases, particularly in terms of gender or age distribution, which may affect the generalizability of the model to diverse patient populations.

Despite these limitations, the dataset remains a valuable resource for developing machine learning models aimed at predicting liver disease [17]. It provides insights into the most critical markers for liver health and allows for an in-depth analysis of how different machine learning algorithms handle these markers. As we will discuss in later sections, the dataset's features play a vital role in the classification accuracy of the algorithms and influence how well they can distinguish between patients with and without liver disease.

3.2 Data Processing

In this test, Rapid Miner with a 10-fold cross-validation operator was used to obtain high accuracy results for each algorithm tested using the dataset [18]. The following is the preprocessing process, as shown in **Figure 3-6**.

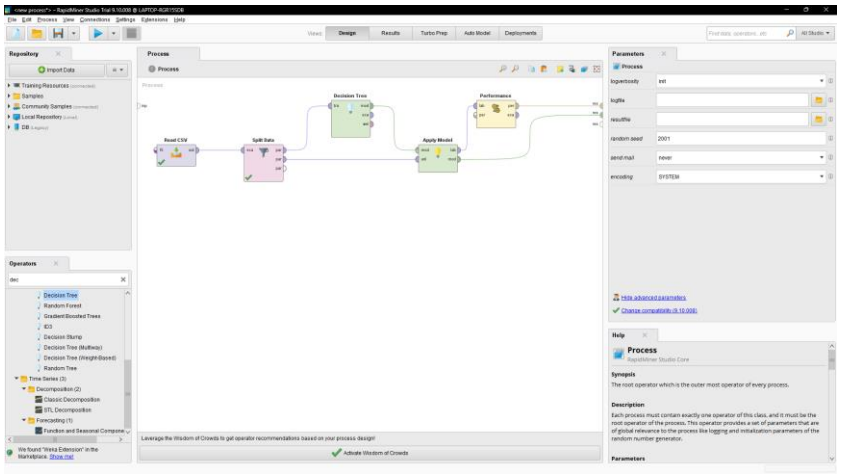

Figure 2. Decision Tree Preprocessing

The first operator is Read CSV, which reads data from a CSV file and imports the dataset into the RapidMiner environment. This operator serves as the starting point for the process flow, providing the raw data for further processing.

After the data is loaded, the Split Data operator is used to divide the dataset into two parts for a training set and a testing set. This division is essential so the model can be trained on one part of the data and tested on the other to objectively measure its performance. Next, the algorithm (decision tree, random forest, naive bayes and k-nearest neighbors) operator builds a classification model based on the training data. This algorithm will generate a structure that represents the classification rules for the attributes in the dataset.

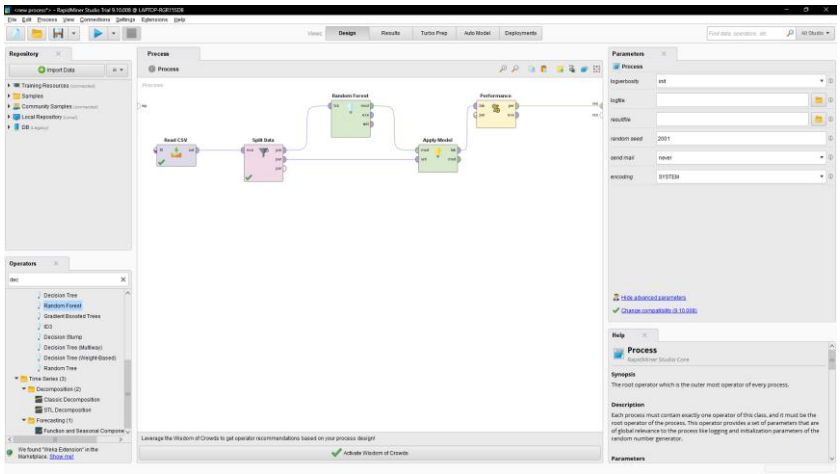

Figure 3. Random Forest Preprocessing

The model results from the algorithm are then connected to the Apply Model operator, which applies the model to the test data to generate predictions. Finally, the Performance operator is used to evaluate the prediction results by calculating various metrics such as accuracy, precision, and recall. This series of operators forms a complete workflow, from data acquisition, splitting, model training, deployment, and performance evaluation.

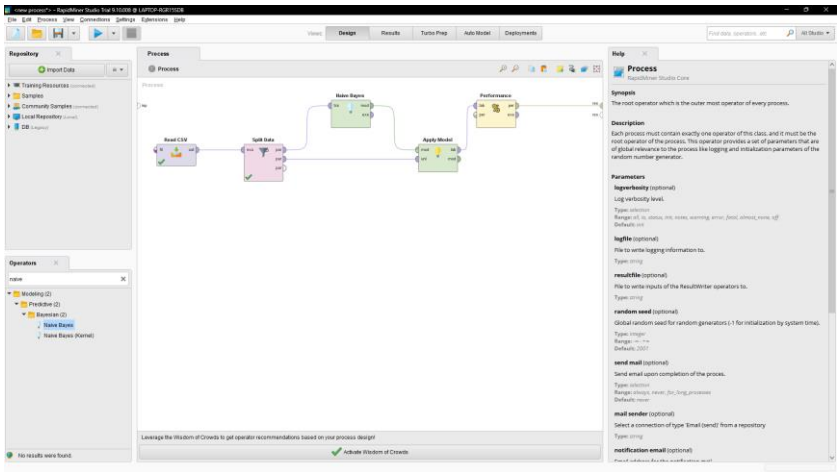

Figure 4. Naive Bayes Preprocessing

3.3 Research Measurement

Figure 7 shows the results of the evaluation of the classification model performance using the Decision Tree algorithm with an accuracy level of 72.41%. From the table, it can be seen that the model successfully classified 81 patients with liver disease correctly, but there were 30 patients with liver disease that were incorrectly classified as not having liver disease. Meanwhile, only 3 patients without liver disease were successfully predicted correctly, and 2 patients without liver disease were actually detected as having liver disease. The precision value for the class of patients with liver disease was 72.97%, while for patients without liver disease it was 60.00%.

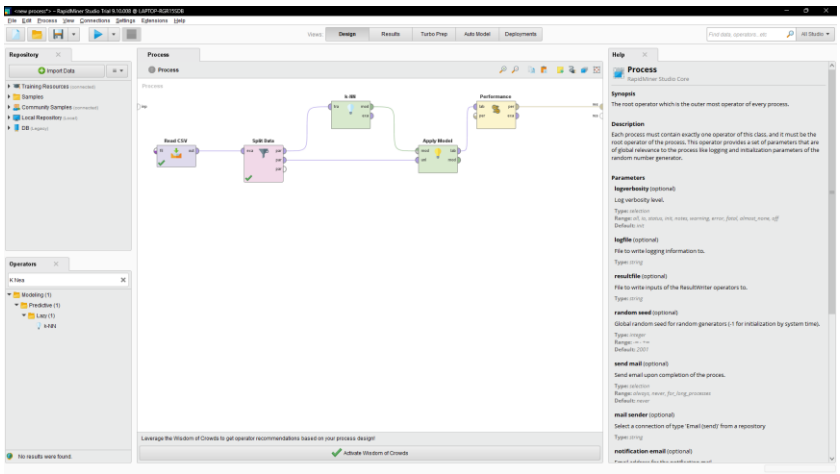

Figure 5. K-Nearest Neighbors Preprocessing

The recall value indicates that the model has excellent ability to detect patients with liver disease, at 97.59%. However, its ability to detect patients without liver disease is still low, at only 9.09%. This means the model is more sensitive to positive cases (patients with liver disease), but less effective in recognizing negative cases (healthy patients). Therefore, although the model has quite good accuracy overall, improvements are still needed in the prediction balance between classes to optimize model performance, especially in reducing detection errors in patients without liver disease.

accuracy: 72.41%

	true Pasien dengan penyakit hati	true Pasien tanpa penyakit hati	class precision
pred. Pasien dengan penyakit hati	81	30	72.97%
pred. Pasien tanpa penyakit hati	2	3	60.00%
class recall	97.59%	9.09%	

Figure 6. Decision Tree Test Result

Figure 8 shows the structure of a decision tree used to classify patients with and without liver disease based on several medical attributes. The first attribute, which serves as the root node, is Age, meaning that patient age is the most influential initial factor in determining the classification results. From this node, branches are formed based on certain threshold values for example, patients with an age below 6.8 are likely to be classified as having no liver disease, while patients with an age above that value are further analyzed based on other attributes such as Alamine Aminotransferase (ALAT), Direct Bilirubin, and Albumin_and_Globulin_Ratio.

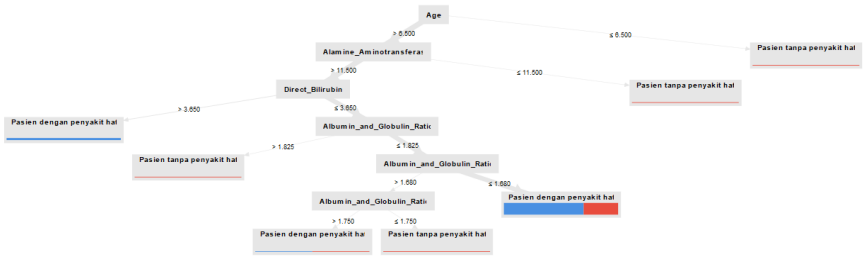

Figure 7. Decision Tree Model

Figure 9 shows the results of the classification model evaluation with an accuracy rate of 72.41% in predicting patients with and without liver disease. Based on the table, the model successfully predicted 81 patients with liver disease correctly, but 30 patients with liver disease were incorrectly classified as not having liver disease. Conversely, there were 3 patients without liver disease that were successfully predicted correctly, and 2 patients without liver disease were incorrectly classified as having liver disease. The precision value for the class of patients with liver disease reached 72.97%, while for patients without liver disease it was 60.00%.

accuracy: 72.41%

	true Pasien dengan penyakit hati	true Pasien tanpa penyakit hati	class precision
pred. Pasien dengan penyakit hati	81	30	72.97%
pred. Pasien tanpa penyakit hati	2	3	60.00%
class recall	97.59%	9.09%	

Figure 8. Random Forest Test Result

Furthermore, the recall values showed a significant difference between the two classes. For patients with liver disease, the recall value reached 97.59%, indicating the model was highly sensitive in detecting positive cases, or truly ill patients. However, the recall for patients without liver disease was only 9.09%, indicating that

the model still struggled to recognize healthy patients. Therefore, although the model is quite good at detecting patients with liver disease, the balance between positive and negative detections still needs to be improved to achieve more accurate and proportional classification results.

accuracy: 60.34%

	true Pasien dengan penyakit hati	true Pasien tanpa penyakit hati	class precision
pred. Pasien dengan penyakit hati	38	1	97.44%
pred. Pasien tanpa penyakit hati	45	32	41.56%
class recall	45.78%	96.97%	

Figure 9. Naive Bayes Test Result

Figure 10 shows the results of the classification model performance evaluation with an accuracy rate of 60.34%. Based on the table, the model successfully predicted 38 patients with liver disease correctly, but there were 45 patients with liver disease that were incorrectly classified as not having liver disease. Conversely, from the group of patients without liver disease, 32 patients were successfully predicted correctly, while only 1 patient without liver disease was incorrectly classified as having liver disease. The precision value for the class of patients with liver disease reached 97.44%, indicating that almost all positive predictions were correct, while for patients without liver disease, the precision value was only 41.56%, indicating a higher level of prediction error.

accuracy: 70.69%

	true Pasien dengan penyakit hati	true Pasien tanpa penyakit hati	class precision
pred. Pasien dengan penyakit hati	72	23	75.79%
pred. Pasien tanpa penyakit hati	11	10	47.62%
class recall	86.75%	30.30%	

Figure 10. K-Nearest Neighbors

In terms of recall, the model showed a significant difference between the two classes. The recall value for patients with liver disease was only 45.78%, meaning the model failed to recognize more than half of the actual positive cases. However, recall for patients without liver disease was very high, at 96.97%, indicating the model was much more reliable in detecting healthy patients. Overall, the model was better at recognizing patients without liver disease than patients with actual liver disease. This indicates that the model is still unbalanced and requires improvements, such as parameter adjustments or the use of other methods, to achieve a more proportional recognition of both classes.

Figure 11 shows the results of the classification model evaluation with an accuracy rate of 70.69%. Based on the table, the model successfully predicted 72 patients with liver disease correctly, but there were 23 patients with liver disease

that were incorrectly classified as not having liver disease. On the other hand, the model also correctly predicted 10 patients without liver disease, but there were still 11 patients without liver disease that were incorrectly classified as having liver disease. The precision value for the class of patients with liver disease reached 75.79%, while for patients without liver disease it was 47.62%, indicating that predictions in patients with liver disease were more accurate than in healthy patients.

Table 2. Overall Test Result

Methods	Accuracy	Precision	Recall
Decision Tree	72.41%	60.00% (positive class: Patients without liver disease)	9.09% (positive class: Patients without liver disease)
Random Forest	72.41%	60.00% (positive class: Patients without liver disease)	9.09% (positive class: Patients without liver disease)
Naïve Bayes	60.34%	41.56% (positive class: Patients without liver disease)	96.97% (positive class: Patients without liver disease)
K-Nearest Neighbors	70.69%	47.62% (positive class: Patients without liver disease)	30.30% (positive class: Patients without liver disease)

In terms of recall, as seen in **Table 2**, the model showed a quite striking difference between the two classes. The recall value for patients with liver disease reached 86.75%, indicating that the model was quite good at detecting truly sick patients. However, the recall for patients without liver disease was only 30.30%, indicating that the model's ability to recognize healthy patients was still low. Overall, although the model had a fairly high level of accuracy, there was still an imbalance in performance between classes. This indicates that the model is more sensitive to positive cases (patients with liver disease) than to negative cases, requiring further adjustments or optimization to achieve more balanced and reliable classification results.

3.4 Analysis of Research Result

The data analysis results demonstrate the performance of several classification algorithms used in the study, specifically Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN). Both the Decision Tree and Random Forest algorithms achieved an accuracy rate of 72.41%, based on testing results from the Rapid Miner application. These algorithms also showed relatively good precision and recall, indicating their ability to effectively identify both patients with liver

disease and those without. This suggests that Decision Tree and Random Forest are suitable for classifying liver disease accurately based on the given dataset.

In contrast, the Naive Bayes algorithm displayed a lower accuracy rate of 60.34%, which was lower than that of Decision Tree and Random Forest. Despite this, Naive Bayes exhibited a high recall rate of 96.97%, which indicates its strong ability to detect patients who actually have liver disease. However, its lower precision, meaning it predicted some healthy patients as having liver disease, highlights the trade-off between recall and precision. This suggests that while Naive Bayes is effective at identifying positive cases, its false positive rate could be a concern in clinical practice.

The K-Nearest Neighbors (KNN) algorithm achieved an accuracy rate of 70.69%, which is slightly lower than both Decision Tree and Random Forest but higher than Naive Bayes. Despite its slightly lower accuracy, KNN still proved to be competitive in predicting liver disease, making it a valuable tool in classification tasks. Overall, while Decision Tree and Random Forest performed best in terms of accuracy, Naive Bayes and KNN each have their advantages, such as higher recall in detecting liver disease, making them viable alternatives depending on the specific needs of a clinical diagnostic system.

3.5 Axiological Perspective and Ethical Considerations in Machine Learning for Medical Diagnoses

The application of machine learning (ML) in medical diagnostics offers significant advantages in terms of accuracy, efficiency, and predictive power [19][20]. However, as with any technology, the integration of ML in healthcare requires careful consideration of its ethical implications. The axiological perspective, which focuses on the study of values, plays a crucial role in understanding the broader implications of using ML in medical settings. Beyond its technical effectiveness, it is essential to assess how these technologies align with fundamental ethical principles, such as fairness, transparency, accountability, and social responsibility.

One of the most critical ethical considerations when using machine learning in medical diagnoses is ensuring fairness [21]. Biases inherent in medical datasets—such as those related to age, gender, or socioeconomic background—can influence the predictions made by ML models. For instance, if a dataset overrepresents a particular demographic (e.g., a specific gender or age group), the trained algorithm may be less accurate in diagnosing individuals from underrepresented groups. This could lead to disparities in healthcare outcomes, particularly if the system is used for clinical decision-making.

To ensure fairness, it is vital to carefully curate datasets, ensuring that they are diverse and representative of the population [22]. Additionally, fairness-aware ML

techniques can be employed to mitigate biases and make the models more equitable [23]. In the context of liver disease diagnosis, for example, it is essential to account for the fact that some medical conditions may present differently across genders or age groups, requiring models that are capable of handling these variations without favoring one group over another.

Another fundamental ethical issue in the use of machine learning for medical diagnoses is the transparency and interpretability of the models. Many machine learning models, particularly deep learning models, are often described as "black boxes," meaning that their decision-making process is not easily understood by humans. In healthcare, where decisions can have life-altering consequences, it is essential that both medical practitioners and patients understand how a model arrives at its predictions.

For example, if a machine learning model diagnoses a patient with liver disease, it is critical for clinicians to understand the reasoning behind the diagnosis [24]. This transparency enables healthcare providers to make informed decisions based on the model's output and allows for accountability in cases where a diagnosis may be incorrect. In the context of liver disease, interpretable models like decision trees or random forests, which provide clear decision paths, can be particularly valuable because they help clinicians understand which features (such as bilirubin levels or enzyme activity) contributed to the diagnosis.

With the increasing use of machine learning in medical settings, questions regarding accountability and liability also emerge [25]. Who is responsible if a machine learning model provides an incorrect diagnosis that leads to adverse health outcomes? Is it the responsibility of the developers who built the algorithm, the healthcare providers who used the system, or both?

Establishing clear lines of accountability is essential in ensuring that patients are protected [26]. Ethical frameworks must be developed to guide the use of ML in healthcare and to determine liability in cases where the model's predictions fail. While machine learning models can assist in diagnosing diseases such as liver dysfunction, human oversight should always be a key component of the diagnostic process, with healthcare professionals remaining responsible for the final decisions based on the model's recommendations.

In any medical application, patient privacy and data security are of paramount importance [27]. Machine learning models require large amounts of data to train effectively, and in healthcare, this data often includes sensitive personal information. Ensuring that patient data is securely handled, anonymized, and stored is crucial to maintaining trust in healthcare systems that utilize ML. Additionally, the use of machine learning models must comply with data protection regulations, such as the

General Data Protection Regulation (GDPR) in Europe, to prevent unauthorized access and misuse of personal health data [28].

Moreover, as the use of artificial intelligence and machine learning in healthcare becomes more widespread, it is essential to establish robust ethical guidelines on how patient data is used, shared, and stored to protect individuals' privacy rights and prevent potential harm [29].

While machine learning models have the potential to improve the accuracy and speed of diagnoses, they should never replace human judgment entirely. Medical professionals must retain the final decision-making authority, particularly when it comes to life-threatening diseases such as liver disease. ML models can support clinicians by providing valuable insights and recommendations, but human expertise is essential in interpreting these recommendations in the context of each patient's unique medical history.

Furthermore, human oversight ensures that ethical considerations are not overlooked when implementing machine learning technologies [30]. Healthcare providers can help mitigate the risks of over-reliance on technology by ensuring that the final clinical decisions are made through a collaborative process that involves both the algorithm's outputs and human experience.

Lastly, the application of machine learning in healthcare must be guided by social and humanitarian values [31]. These technologies should be developed and deployed with the primary goal of improving patient outcomes and advancing public health, rather than for profit or technological advancement alone. The deployment of ML systems should aim to enhance access to quality healthcare for all patients, particularly in underserved or low-resource settings.

In the context of liver disease, for example, ML systems can help diagnose conditions more quickly and accurately, potentially reducing healthcare costs and improving access to treatment. However, ethical considerations must be integrated into the design and deployment of these systems to ensure that they contribute positively to the well-being of patients and society at large.

4. Conclusion

This study evaluates the effectiveness of four machine learning algorithms—Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN)—in diagnosing liver disease. Decision Tree and Random Forest both achieved 72.41% accuracy, showing strong performance but with some limitations in classifying healthy patients. Naïve Bayes, with a lower accuracy of 60.34%, excelled in recall (96.97%), making it effective at detecting liver disease, despite precision errors. KNN, with 70.69% accuracy, proved to be a competitive alternative. Beyond technical performance, the study addresses the ethical implications of using machine

learning in healthcare. It emphasizes the importance of fairness, transparency, and human oversight to ensure that algorithms are used responsibly and complement clinical judgment. The research highlights the societal responsibility of deploying machine learning systems in healthcare, urging the integration of ethical standards to improve diagnostic accuracy and accessibility, especially in underserved communities.

Authors' Declaration

Authors' contributions and responsibilities - The authors made substantial contributions to the conception and design of the study. The authors took responsibility for data analysis, interpretation, and discussion of results. The authors read and approved the final manuscript.

Funding - No funding information from the authors.

Availability of data and materials - All data is available from the authors.

Competing interests - The authors declare no competing interest.

Additional information - No additional information from the authors.

References

[1] W. Atifi et al., "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," [Online]. Available: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0330899&type=printable

[2] S. Velu, V. Ravi, and K. Tabianan, "Data mining in predicting liver patients using classification model," [Online]. Available: https://link.springer.com/content/pdf/10.1007/s12553-022-00713-3.pdf

[3] [Missing authors], "[PDF] PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK PREDIKSI ...," [Online]. Available: https://jurnal.bsi.ac.id/index.php/reputasi/article/download/109/37

[4] N. Nia, E. Kaplanoğlu, and A. Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction," [Online]. Available: https://link.springer.com/content/pdf/10.1007/s44163-023-00049-5.pdf

[5] A. Sultana and R. Islam, "Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification," [Online]. Available: https://jesit.springeropen.com/counter/pdf/10.1186/s43067-023-00101-5

[6] [Missing authors], "Applying Naive Bayesian Networks to Disease Prediction - PMC - NIH," [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5203736/

[7] [Missing authors], "[PDF] PREDICTION OF LIVER DISEASE WITH RANDOM FOREST ...," [Online]. Available: https://ijrdst.org/public/uploads/paper/842421702539946.pdf

[8] M. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of Heart Disease

Using K- Nearest Neighbor and Genetic Algorithm," [Online]. Available: https://arxiv.org/pdf/1508.02061

[9] [Missing authors], "Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity," [Online]. Available: https://escholarship.org/uc/item/58n1z2hr

[10] H. Bharadwaj et al., "Artificial Intelligence in Population-Level Gastroenterology and Hepatology: A Comprehensive Review of Public Health Applications and Quantitative Impact," [Online]. Available: https://doi.org/10.1007/s10620-025-09452-7

[11] [Missing authors], "Axiology and the Evolution of Ethics in the Age of AI: Integrating Ethical Theories via Multiple-Criteria Decision Analysis †," [Online]. Available: https://www.mdpi.com/2504-3900/126/1/17

[12] [Missing authors], "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke ...," [Online]. Available: https://www.ojs.stmikplk.ac.id/index.php/saintekom/article/view/352

[13] [Missing authors], "ILPD (Indian Liver Patient Dataset) Data Set - Kaggle," [Online]. Available: https://www.kaggle.com/datasets/rahulrajpandey31/ilpd-indian-liver-patient-dataset-data-set

[14] [Missing authors], "Comparative Analysis of Machine Learning Algorithms : Random Forest algorithm, Naive Bayes Classifier and KNN - A survey," [Online]. Available: https://jrps.shodhsagar.com/index.php/j/article/view/556

[15] K. Sujon et al., "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," [Online]. Available: https://link.springer.com/content/pdf/10.1186/s40537-025-01313-4.pdf

[16] [Missing authors], "ILPD (Indian Liver Patient Dataset)," [Online]. Available: https://datasets.aim-ahead.net/dataset/p/UCI_DS_225

[17] S. Ganie, P. K. Pramanik, and Z. Zhao, "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches," [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/counter/pdf/10.1186/s12911-024-02550-y

[18] [Missing authors], "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi ...," [Online]. Available: https://jsi.politala.ac.id/index.php/JSI/article/view/622

[19] [Missing authors], "Ilmu dalam Tinjauan Filsafat: Ontologi, Epistemologi, dan Aksiologi," [Online]. Available: https://ejournal.stai-tbh.ac.id/al-aulia/article/view/1875

[20] [Missing authors], "Sistematika Filsafat Menurut Ontologi, Epistemologi, dan Aksiologi ...," [Online]. Available: https://www.researchgate.net/publication/372771565_Sistematika_Filsafat_Menurut_Ontologi_Epistemologi_dan_Aksiologi_dalam_Artificial_Intelligence

[21] D. Ueda et al., "Fairness of artificial intelligence in healthcare: review and recommendations," [Online]. Available:

https://link.springer.com/content/pdf/10.1007/s11604-023-01474-3.pdf
[22] N. Shahbazi et al., "Representation Bias in Data: A Survey on Identification and Resolution Techniques," [Online]. Available: https://doi.org/10.1145/3588433
[23] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3616865
[24] R. Agrawal et al., "Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," [Online]. Available: https://diagnosticpathology.biomedcentral.com/counter/pdf/10.1186/s13000-025-01686-3
[25] M. Mello and N. Guha, "Understanding Liability Risk from Using Health Care Artificial Intelligence Tools," [Online]. Available: https://doi.org/10.1056/nejmhle2308901
[26] E. Aveling, M. Parker, and M. Dixon-Woods, "What is the role of individual accountability in patient safety? A multi-site ethnographic study," [Online]. Available: https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/1467-9566.12370
[27] L. Jawad, "Security and Privacy in Digital Healthcare Systems: Challenges and Mitigation Strategies," [Online]. Available: https://doi.org/10.1177/09702385241233073
[28] J. Starkbaum and U. Felt, "Negotiating the reuse of health-data: Research, Big Data, and the European General Data Protection Regulation," [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/2053951719862594
[29] S. Yu, S. Lee, and H. Hwang, "The ethics of using artificial intelligence in medical research," [Online]. Available: https://doi.org/10.7180/kmj.24.140
[30] Y. Bengio et al., "International Scientific Report on the Safety of Advanced AI (Interim Report)," [Online]. Available: https://arxiv.org/pdf/2412.05282
[31] I. Chen et al., "Ethical Machine Learning in Healthcare," [Online]. Available: https://arxiv.org/pdf/2009.10576