Comparison of Machine Learning Algorithms for Liver Disease Classification

*Nur Futri Ayu Jelita
Department of Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia
12250320374@students.uin-suska.ac.id

Nayla Husna Ryanda
Department of Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia
12250321428@students.uin-suska.ac.id

Fatimah Fhingkan Agustina
Department of Information Systems, Faculty of Computer Science, Universitas Muhammadiyah Riau, Indonesia
220402123@student.umri.ac.id

Abstract— *Liver disease is a serious medical condition often diagnosed late due to the lack of early symptoms and limited access to fast and affordable diagnostic technologies. This study utilized the Indian Liver Patient Dataset (ILPD) to develop a machine learning-based liver disease prediction model. The research process included data collection, preprocessing, model building, and performance comparison of various algorithms, such as Naive Bayes, K-Nearest Neighbor (KNN), Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Network. The evaluation results revealed that Logistic Regression achieved the best performance with an accuracy of 72.00%, precision of 91.80%, and recall of 74.70%, offering a balance between accurate detection and minimal diagnostic errors. This study concludes that Logistic Regression is the most effective algorithm for liver disease prediction, supporting early detection and medical decision-making.*

Keywords— *Liver disease, classification, machine learning, Random Forest, ILPD dataset*

I. INTRODUCTION

Liver disease is a dangerous medical condition that can be fatal if not treated promptly. The liver does many things for the body, such as metabolising nutrients, storing energy, and detoxifying toxic substances. Disorders in the liver can affect metabolism and can lead to serious problems affecting other organs. Unfortunately, many cases of liver disease are only discovered at an advanced stage as early-stage disease usually does not show any symptoms. Medical personnel face many difficulties when it comes to early detection, which is crucial for reducing the risk of complications and improving the patient's prognosis [1].

One of the main obstacles in detecting liver disease is the lack of access to fast, accurate and inexpensive diagnostic technologies. Diagnosis usually takes a long time and blood tests are expensive to determine certain liver conditions [2]. Due to these conditions, patients struggle to get adequate diagnosis and treatment, especially those living in areas with limited healthcare resources. In addition, medical personnel face difficulties in detecting liver disease early because they cannot manually analyze clinical data. Without a good decision support system, medical personnel will find it difficult to determine high-risk patients based on existing clinical data [3].

The Indian Liver Patient Dataset (ILPD) offers an opportunity to address this issue. Using machine teaching techniques, predictive models can be built with readily available clinical data to help detect liver disease at an early stage. However, there are issues that arise when choosing the right algorithms to process these datasets to produce accurate and reliable models. This includes proper attribute analysis, good data processing, and selection of algorithms that can deliver accurate and optimized results in a time-efficient manner. [4].

To solve this problem, the Indian liver disease (ILPD) patient dataset can be used to build a liver disease prediction model using machine teaching methods. A problem has

to be solved in several stages. Firstly, clinical characteristics such as age, gender, bilirubin levels, liver enzymes, and protein levels in the blood associated with liver disease detection are incorporated into the data structure of the ILPD dataset before data collection and pre-processing. Firstly, the data is cleaned of empty values, standardised or normalized if necessary, and converted into numerical data so that machine learning algorithms can process it. This step is crucial to ensure that the data used is clean and ready for model training [4].

Next, the data is examined to see the distribution of each feature and its relationship with liver disease status. The purpose of this process is to find the features that have the most impact on liver disease classification, so feature selection techniques can be used to select the most relevant features and improve model accuracy. To predict liver disease status, various machine learning algorithms were tested and compared. Appropriate machine learning algorithms were selected using algorithms such as Naive Bayes, K-Nearest Neighbor (KNN), C4.5, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Network [5].

Next comes model training and validation. Here, the model is trained with training data and evaluated with validation data. A cross-validation method will be used to ensure that the model's training data is not larger than the actual. After training, its performance is assessed using metrics such as accuracy, precision, and recall to identify the best performing model.

Using the patient's clinical data, the model can be used to make early predictions about the risk of liver disease. Hopefully, the system will advise doctors to perform additional tests on high-risk patients. In addition, this prediction system can be integrated into existing electronic medical record systems in hospitals or clinics. Before the system is widely used, field testing should be conducted to ensure that the model works well in a real environment with feedback from end users such as doctors and medical personnel. Finally, the model should be updated regularly to keep up with changes in data and machine learning algorithms for the system to be used effectively [6].

II. RESEARCH METHODOLOGY

Overall, this research included 5 stages. The first stage is data collection. The second stage is data preprocessing. The third stage is the division of data into training data and test data. [7]. The fourth stage is model building. The last stage is performance comparison. All stages of this research are shown in Figure 1.

A. Data Collection

The Indian liver patient dataset (ILPD), downloaded from the UCI Machine Learning Repository platform, contains information on various factors related to liver health, such as age, gender, bilirubin levels, liver enzymes, and the patient's final diagnosis status, whether they have liver disease or not.

B. Preprocess Data

Orange helps prepare the dataset before further analysis during the data preprocessing stage. Firstly, the feature values are normalised into the interval [0,1] in Orange's preprocessing efforts to ensure that the values of each attribute in the dataset have the same scale. This improves the performance of the classification model used in this study [8].

C. Division of Training Data and Test Data

The K-Fold Cross Validation method divides the test and training data into ten parts. At each iteration, one part is used as test data and the other nine parts as training data;

this process is repeated ten times, so that each part of the data is used only once as test data. This method helps to divide the data into training and testing data, and estimate the error in predicting model performance [9].

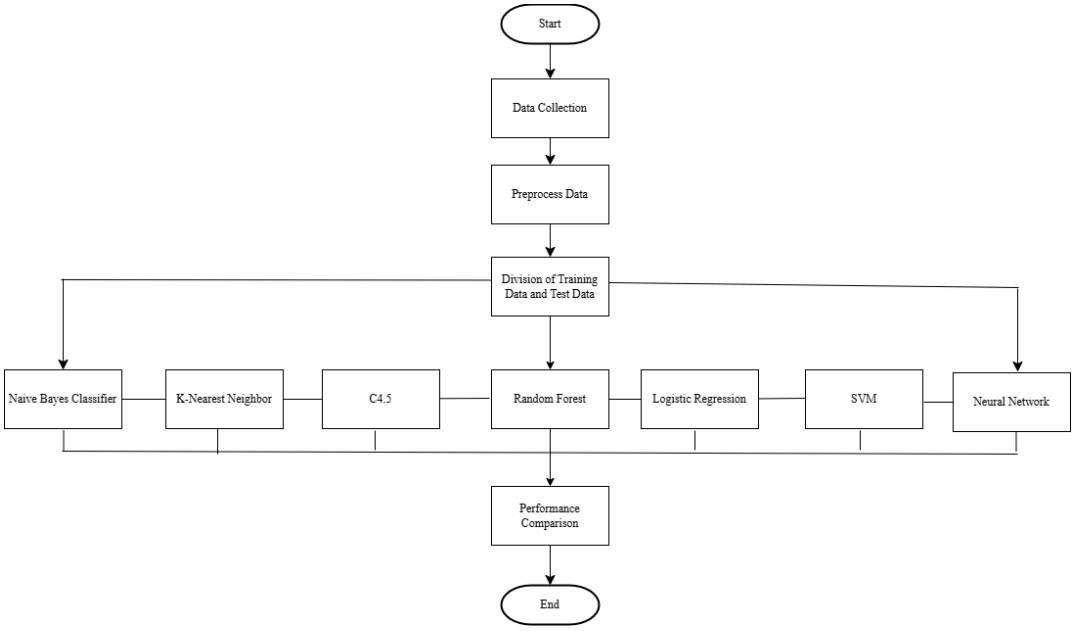

Figure 1. Research Methodology

D. Classification Modelling

In this study, 7 classification algorithms were used, namely Naive Bayes, K-Nearest Neighbor (KNN), C4.5, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Network.

1) Naïve Bayes Classifier

In the NBC algorithm, the Naive Bayes (NBC) operator is used in the Orange tool. The Naïve Bayes method is a statistical approach to perform induction inference on classification problems [10].

2) K-Nearest Neighbor

In the K-NN algorithm, 10 trials were conducted by trying various values of the K parameter. The values of K tried in K-NN were K = 3, 5, 7, 9, 11, 13, 15, 17, 19, and 21 [11].

3) C4.5

In the C4.5 algorithm, the operator used is the Decision Tree operator in the model column in the Orange tools. The Decision Tree algorithm can be used to predict or classify an event by forming a decision tree [12].

4) Random Forest

Random Forest is a combination of each existing decision tree techniques, which are then combined and integrated into a single model. It is used to improve prediction accuracy by building multiple decision trees [13].

5) Logistic Regression

Used to predict the probability of data belonging to one of two classes (binary), although it can also be used for multi-class cases [14].

6) Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm in data mining used for classification, regression, pattern recognition, dimension reduction, and anomaly detection [15]. SVM can also classify data into two or more categories and has the ability to predict continuous values through Support Vector Regression (SVR) [16]. Applications such as facial recognition, email spam classification, text analysis, and bioinformatics often use SVM to find outliers in datasets. The advantage of SVM is its ability to handle complex and high-dimensional data [9].

7) Neural Network

Neural networks are data mining algorithms based on how the human brain works. They are also used for classification, regression, and pattern recognition. Image recognition, sound analysis, natural language processing, and predictive tasks utilise their ability to find complex relationships in data, making them a common choice [17]. Due to their ability to learn from large and complex datasets, neural networks are powerful tools in a variety of applications, such as product recommendations and autonomous vehicles [18].

E. Performance Comparison

Each classification model was tested using a confusion matrix to measure performance, including accuracy, precision, and recall. This was done using the performance operator in the Orange tool. A comparison was made between seven algorithm models: Naive Bayes, K-Nearest Neighbour (KNN), C4.5, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Network. These algorithmic models use K-NN analysis based on variations in the K parameter value. Equations 2 and 3 show the formulas for precision and recall.

$$\text{Accurasy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{x } 100\% \quad (1)$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{x } 100\% \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{x } 100\% \quad (3)$$

III. RESULTS AND DISCUSSION

In the results section, the dataset was processed through several main stages, starting with data collection, followed by data preprocessing to ensure the quality of the data used. After that, the data was divided into training data and test data. The process continues with modelling using seven classification algorithms, namely Naive Bayes, K-Nearest Neighbour (KNN), C4.5, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Network. Each model that has been built is then evaluated for performance through a performance comparison process to determine the best model for predicting the test data.

A. Data Collection Results

At this point, the search was conducted using the Indian Liver Patient Dataset (ILPD), which was downloaded from the UCI Machine Learning Repository. This dataset contains information about various risk factors for liver disease in Microsoft Excel (.xlsx) format, with a total of 582 data points consisting of eleven attributes that determine risk factors for liver disease. The Indian Liver Patient (ILPD) dataset has data attributes

described in Table 1. Table 2 displays a sample of data from potential patients predicted to have liver disease.

Table 1. Explanation of Data Attributes

Attributes	Code	Explanation	Description
Age	A1	Patient Age	Age in years
Gender	A2	Patient gender	Male, Female
Total_Bilirubin	A3	Total bilirubin in the blood	Total bilirubin level (0.1 mg/dl) in the blood. An increase in this level may indicate liver problems.
Direct_Bilirubin	A4	Direct bilirubin in the blood	Direct bilirubin level (0.7 mg/dl), part of total bilirubin. Often elevated in liver disorders.
Alkaline_Phosphotase	A5	Alkaline phosphatase enzyme in the blood	Alkaline phosphatase enzyme level (187 IU/L). Important for liver and bone metabolism.
Alamine_Aminotransferase (ALT)	A6	Alamine aminotransferase enzyme in the blood	ALT level (16 IU/L). An enzyme released into the blood when liver damage occurs.
Aspartate_Aminotransferase (AST)	A7	Aspartate aminotransferase enzyme in the blood	AST level (18 IU/L). This enzyme increases if there is damage to the liver or muscles.
Total_Proteins	A8	Total protein in the blood	Total protein in the blood (6.8 g/dL) is important for the growth and repair of body tissues.
Albumin	A9	Albumin levels in the blood	Albumin level in blood, (3.3 g/dL). The main protein in blood produced by the liver.
Albumin_and_Globulin_Ratio	A10	Albumin to globulin ratio in blood	The ratio between albumin and globulin (0.9 A/G) which describes the condition of liver health.
Dataset	A11	Label target whether the patient has liver disease or not	1 indicates the presence of liver disease, and 2 indicates the absence of liver disease.

Table 2. Patient Data Sample

No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
P1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
P2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
P3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
P4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
P5	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
P6	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1
P7	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.10	1
P8	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.20	2
P9	55	Male	0.7	0.1	290	53	58	6.8	3.4	1.00	1
P10	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.80	1
.....	....	.....	....	....	.....	.....	.....	.....	.....	.....	.....
P581	31	Male	1.3	0.5	184	29	32	6.8	3.4	1.00	1
P582	38	Male	1.0	0.3	216	21	24	7.3	4.4	1.50	2

B. Pre-processed Data Results

The initial patient data consisted of 582 patients, and in the next stage, 414 patients were detected with liver disease. After obtaining the data on patients detected with liver disease, the next step was min-max normalisation. The normalised data sample can be seen in Table 3.

Table 3. Data Normalization Sample

No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
P1	0.6744	Male	0.14075	0.27551	0.310699	0.027136	0.018296	0.6957	0.50	0.176	1
P2	0.6744	Male	0.09249	0.20408	0.208598	0.025126	0.011791	0.6232	0.5217	0.236	1
P3	0.6279	Male	0.00804	0.01531	0.058134	0.002010	0.002033	0.5942	0.5434	0.280	1
P4	0.7907	Male	0.04692	0.09694	0.064485	0.008543	0.009961	0.6667	0.3261	0.040	1
P5	0.4884	Male	0.01877	0.03061	0.070835	0.004523	0.000813	0.7101	0.7609	0.400	1
P6	0.2558	Female	0.0670	0.00510	0.044455	0.003015	0.000407	0.6232	0.5652	0.280	1
P7	0.2907	Female	0.0670	0.01020	0.067904	0.002010	0.000203	0.5797	0.5870	0.320	1
P8	0.1512	Male	0.0670	0.01020	0.067904	0.006030	0.001830	0.6812	0.6957	0.360	2
P9	0.5930	Male	0.00402	0.00510	0.110894	0.021608	0.009758	0.5942	0.5435	0.280	1
P10	0.6163	Male	0.00268	0.000	0.071812	0.020603	0.009961	0.4638	0.3913	0.200	1
.....	....	.....	....	....	.....	.....	.....	.....	.....	.....	.....
P581	0.3140	Male	0.01206	0.2041	0.059111	0.009548	0.004472	0.5942	0.5435	0.280	1
P582	0.3953	Male	0.00804	0.01020	0.074744	0.005528	0.002846	0.6667	0.7609	0.480	2

C. Results of Classification Model Creation

Based on the experimental results, the NBC algorithm produced an accuracy value of 66.33%, a precision value of 65.06%, and a recall value of 84.11%, while the K-NN algorithm (K=19) produced an accuracy value of 70.30%, a precision value of 66.20%, and a recall value of 70.3%. The experimental results show that the NBC algorithm performs quite well in terms of accuracy, but it also has shortcomings in terms of precision. Table 4 shows the NBC confusion matrix, Table 5 shows the K-NN confusion matrix, Table 6 shows the C4.5 confusion matrix, and Table 7 shows the Random Forest confusion matrix. Table 8 shows the confusion matrix for logistic regression, Table 9 shows the confusion matrix for SVM, and Table 10 shows the confusion matrix for Neural Network.

Table 4. Confusion Matrix NBC

		Original Class	
		True 1	**True 2**
Pred	**Yes**	270	145
Pred	**No**	51	116

Table 5. Confusion Matrix K-NN

		Original Class	
		True 1	**True 2**

Pred	**Yes**	385	30
Pred	**No**	138	39

Table 6. Confusion Matrix C4.5

		Original Class	
		True 1	**True 2**
Pred	**Yes**	342	73
Pred	**No**	104	63

Table 7. Confusion Random Forest

		Original Class	
		True 1	**True 2**
Pred	**Yes**	355	60
Pred	**No**	117	50

Table 8. Confusion Logistic Regression

		Original Class	
		True 1	**True 2**
Pred	**Yes**	381	34
Pred	**No**	129	38

Table 9. Confusion SVM

		Original Class	
		True 1	**True 2**
Pred	**Yes**	309	106
Pred	**No**	108	59

Table 10**.** Confusion Neural Network

		Original Class	
		True 1	**True 2**
Pred	**Yes**	372	43
Pred	**No**	126	41

The K-NN experiment was conducted ten times with different K values, namely 3, 5, 7, 9, 11, 13, 15, 17, 19, and 21. The results of the experiment are shown in Table 7. Since the recall value is balanced with accuracy, at 70.3%, K=19 is the most accurate value for this study. This is because the accuracy, precision, and recall values of this study are relatively high compared to the other K values.

Table 11. Accuracy Algorithm K-NN

Value K	**Accuracy Value (%)**	**Precision Value (%)**	**Recall Value (%)**
3	65,6%	63,8%	65,6%
5	66,8%	64,9%	66,8%
7	67,5%	64,6%	67,5%
9	68,4%	65,4%	68,4%
11	68,6%	65,4%	68,6%
13	68.9%	65%	68,9%
15	69,2%	65,3%	69,2%
17	69,4%	65,1%	69,4%
19	**70,3%**	**66,2%**	**70,3%**
21	69,9%	65,5%	69,9%

The experiment was conducted with various numbers of decision trees for the Random Forest algorithm, including 50, 100, 200, 500, and 1000. The results of the study can be seen in Table 12. The best accuracy value was 69.9%, the precision value was 66.6%, and the recall value was 69.9%. Therefore, 100 trees were selected because they provided the most balanced results between accuracy, precision, and recall in this study compared to other numbers of trees.

Table 12. Accuracy Algorithm Random Forest

Number Of Trees	Accuracy Value (%)	Precision Value (%)	Recall Value (%)
50	67,7%	64,5%	67,7%
100	**69,9%**	**66,6%**	**69,9%**
200	69.2%	65.8%	69,2%
500	69.4%	66,4%	69.4%
1000	68,2%	64,8%	68,2%

The experiment was conducted on the Logistic Regression algorithm with three types of regularisation: Lasso (L1), Ridge (L2), and no regularisation (None). The results are shown in Table 13. The Lasso (L1) regularisation type showed the best performance with an accuracy value of 72.0%, precision of 68.4%, and recall of 72.0%. Meanwhile, the Ridge (L2) regularisation type showed a slight decrease in performance with an accuracy value of 72.0% and precision of 68.4%.

Table 13. Accuracy Algorithm Logistic Regression

Regularization Type	Accuracy Value (%)	Precision Value (%)	Recall Value (%)
Lasso (L1)	**72,0%**	**68,4%**	**72,0%**
Ridge (L2)	71,6%	67,3%	71,6%
None	70,6%	66,8%	70,6%

The experiment was conducted using various types of kernels for the Support Vector Machine (SVM) algorithm, including RBF, Linear, Polynomial, and Sigmoid kernels. The results showed that the RBF kernel had the best accuracy at 63.2%, followed by precision at 63.2% and recall at 63.2%. The Linear kernel also had good accuracy with 62.9%, precision of 62.9%, and recall of 62.9%. Meanwhile, the Sigmoid kernel had fairly good accuracy with 62.9%, and precision.

Table 14. Accuracy Algorithm SVM

Kernel	Accuracy Value (%)	Precision Value (%)	Recall Value (%)
Linear	49,8%	68,6%	49,8%
Polynomial	44,5%	71,6%	44,5%
RBF	**63,2%**	**63,1%**	**63,2%**
Sigmoid	62,9%	62,9%	62,9%

In the Neural Network algorithm, as shown in Table 15, experiments were conducted using various activation functions and solvers. The best results were obtained with the combination of the Identify activation function and the L-BFGS-B solver, achieving an accuracy of 70.8%, a precision of 67.0%, and a recall of 70.8%. Other combinations, such as the Adam activation function, also yielded similar accuracy, at 71.1% with a precision of 67.3% and a recall of 71.1%. Based on these results, the combination of Identify and L-BFGS-B was selected as the best.

Table 15. Accuracy Algorithm Neural Network

Activation	Solver	Accuracy Value (%)	Precision Value (%)	Recall Value (%)
	L-BFGS-B	**70,8%**	**67,0%**	**70,8%**
Identify	SGD	70,3%	64,2%	70,3%
	Adam	71,1%	67,3%	71,1%
Logistic	L-BFGS-B	68,6%	68,7%	68,6%
	SGD	71,3%	50,8%	71,3%
	Adam	70,8%	67,0%	70,8%
Tanh	L-BFGS-B	68,6%	68,3%	68,6%
	SGD	70,1%	63,9%	70,1%
	Adam	71,0%	67,3%	71,0%
ReLu	L-BFGS-B	67,4%	67,4%	67,4%
	SGD	71,1%	64,7%	71,1%
	Adam	68,6%	67,9%	68,6%

The decision tree generated by the C4.5 algorithm is shown in Figure 2. In this decision tree, it can be seen that only 8 of the 13 attributes are used, namely:

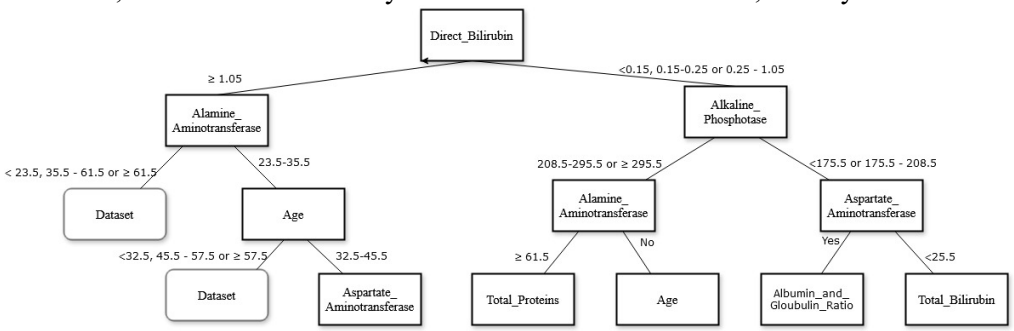

Figure 2. Decision Tree C4.5

D. Performance Comparison

This study tends to choose the logistic regression model as the best algorithm for predicting liver disease, based on existing experimental results. The evaluation results show that the logistic regression model provides the best balance of performance between accuracy, precision, and recall.

Logistic Regression has an accuracy of 72.00% in predicting most of the data. The very high performance (91.80%) ensures that positive predictions are rarely incorrect, making this algorithm highly reliable in identifying patients who truly have liver disease without causing too many misdiagnoses in healthy patients. Additionally, the recall value of 74.70% indicates that this algorithm is sufficiently sensitive in identifying patients who truly have liver disease, although this value is slightly lower than that of the Naive Bayes Classifier.

Considering this performance balance, the logistic regression model is considered the most suitable for application in liver disease prediction scenarios, where the balance between accurate detection and minimal misdiagnosis is a top priority, as shown in Figure 3.

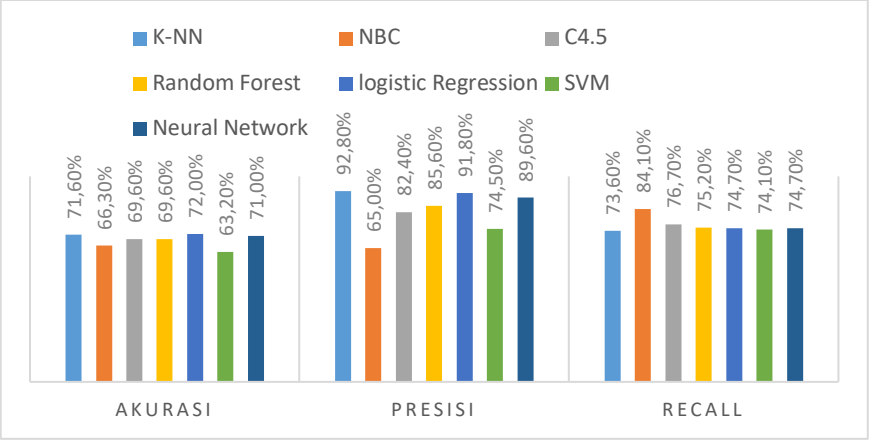

Figure 3. Comparison of Classification Model Performance

IV. Conclusion

Based on the results of the analysis and evaluation of the performance of various classification algorithms, this study found that the Logistic Regression algorithm is the most superior model for predicting liver disease. Logistic Regression achieves the highest accuracy rate of 72.00%, precision of 91.80%, and recall of 74.70%, demonstrating an

excellent balance between accurate prediction capabilities, reliable positive predictions, and effective detection of patients with liver disease. These findings align with M. Mardewi's research, which states that Logistic Regression demonstrates good capability in handling linear relationships between clinical variables and the likelihood of cirrhosis of the liver [19]. In addition, similar results were also found in a study by H. Hikmayanti Handayani et al, which showed that Logistic Regression was able to provide competitive performance for prediction [20].

In addition, the Naive Bayes Classifier (NBC) algorithm showed superiority in detecting patients who were actually diagnosed with liver disease with the highest recall of 84.10%, but had weaknesses in accuracy (66.30%) and precision (65.00%), making it less reliable than Logistic Regression in this context. Other algorithms such as K-NN, Random Forest, C4.5, SVM, and Neural Network also provide fairly good results, but are unable to match the balanced performance demonstrated by Logistic Regression. Among them, K-NN has high precision (92.80%), while Random Forest and Neural Network provide stable results, although slightly lower in accuracy. Thus, this study concludes that Logistic Regression is the most suitable algorithm to be applied in liver disease prediction scenarios, as it is capable of providing accurate, reliable, and balanced results in detecting patients with liver disease.

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] Wiwid Wahyudi, "Implementasi Data Mining Untuk Klasifikasi Penyakit Liver Dengan C4.5 Adaboost," *J. Ilm. Tek. Inform. dan Komun.*, vol. 1, no. 3, pp. 71–76, 2021.

[3] S. Zulaikhah Hariyanti Rukmana, A. Aziz, and W. Harianto, "Optimasi Algoritma K-Nearest Neighbor (Knn) Dengan Normalisasi Dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 439–445, 2022.

[4] I. Setiawati, A. P. Wibowo, and A. Hermawan, "Pendahuluan Tinjauan Pustaka Penelitian Sebelumnya Klasifikasi," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.

[5] V. Wulandari, W. J. Sari, Z. Alfian, L. Legito, and T. Arifianto, "Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 710–718, 2024.

[6] A. D. W. M. Sidik, I. Himawan Kusumah, A. Suryana, Edwinanto, M. Artiyasa, and A. Pradiftha Junfithrana, "Gambaran Umum Metode Klasifikasi Data Mining," *Fidel. J. Tek. Elektro*, vol. 2, no. 2, pp. 34–38, 2020.

[7] N. Utami, K. Ahmad Baihaqi, E. E. Awal, D. Wahiddin, and F. I. Komputer, "Analisis Kinerja Algoritma Decision Tree Dan Random Forest Dalam Klasifikasi Penyakit Kardiovaskular," *Technol. Sci.*, vol. 6, no. 2, pp. 970–980, 2024.

[8] N. Maulida, N. Suarna, and W. Prihartono, "Analisis Ulasan Sentimen Aplikasi Mobile Jkn Dengan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1651–1658, 2024.

[9] N. Widiastuti, A. Hermawan, and D. Avianto, "Implementasi Metode Naïve Bayes Untuk Klasifikasi Data Blogger," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 3, pp. 985–994, 2023.

[10] J. I. Marzuki, K. Mataram, and N. T. Bar, "KOMPARASI AKURASI METODE CORRELATED NAIVE BAYES CLASSIFIER DAN NAIVE BAYES CLASSIFIER UNTUK DIAGNOSIS PENYAKIT DIABETES Hairani , Gibran Satya Nugraha ,

Mokhammad Nurkholis Abdillah , Muhammad Innuddin InfoTekJar (Jurnal Nasional Informatika dan Teknolog," pp. 6–11.
[11] S. Zhang, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp. 4663–4675, 2022.
[12] N. T. Rahman, "Analisa Algoritma Decision Tree Dan Naïve Bayes Pada Pasien Penyakit Liver," *J. Fasilkom*, vol. 10, no. 2, pp. 144–151, 2020.
[13] F. Kurniawan and Ivandari, "Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara," *IC-Tech*, vol. XII, no. 1, pp. 1–8, 2017.
[14] Y. Handayani *et al.*, "PERBANDINGAN ALGORITMA LOGISTIC REGRESSION DAN NAÏVE," vol. 4307, no. May, pp. 1435–1440, 2025.
[15] S. Muawanah, U. Muzayanah, M. G. R. Pandin, M. D. S. Alam, and J. P. N. Trisnaningtyas, "Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia," *Qubahan Acad. J.*, vol. 3, no. 4, pp. 206–218, 2023.
[16] E. Pusporani and S. Qomariyah, "323508-Klasifikasi-Pasien-Penderita-Penyakit-Li-496B23E3," vol. 2, no. March, pp. 25–32, 2019.
[17] A. I. Kushartanto and R. T. Aldisa, "Data Mining Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes dalam Prediksi Penerimaan Beasiswa," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 196–207, 2023.
[18] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–14, 2018.
[19] M. Mardewi, "Prediksi Dini Liver Cirrhosis Untuk Kesehatan Hati Menggunakan Metode Machine Learning.," *Adv. Comput. Syst. Innov. J.*, vol. 1, no. 1, pp. 87–91, 2023.
[20] H. Hikmayanti Handayani, K. Ahmad Baihaqi, and U. Buana Perjuangan Karawang, "Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke," *Syntax J. Inform.*, vol. 12, no. 01, pp. 15–23, 2023.