

Random Search Optimization Using Random Forest Algorithm For Liver Disease Prediction

Riyan Bayu Satriya¹, Kusnawi²

^{1,2}Informatics, Faculty of Computer Science, AMIKOM University, Yogyakarta, Indonesia

Keywords: Liver, Machine Learning, Random Forest, Optimization, Random Search	Abstrak
Submitted: 05/May/2025	<p>The liver is a vital human organ with complex and diverse functions. One of the diseases that affect the liver is hepatitis or liver disease. Early detection is crucial to enable more effective intervention and slow the progression of the disease. However, diagnosing liver disease often faces challenges, especially in detecting the early stages of the disease from complex and diverse medical data. This study aims to optimize the <i>Random Forest</i> algorithm using the <i>Random Search</i> method for liver disease detection. The <i>Random Forest</i> algorithm is applied as the primary model in this research, while hyperparameter optimization is performed using the <i>Random Search</i> method to enhance model performance. The results show that the <i>Random Forest</i> model without optimization achieves an accuracy of 93%. After hyperparameter optimization, the model's accuracy increases to 94%. In conclusion, applying hyperparameter optimization using the <i>Random Search</i> method successfully improves the performance of the <i>Random Forest model</i>. The resulting model provides more accurate predictions.</p>
Revised: 14/May/2025	
Accepted: 20/May/2025	
Corresponding Author: Riyan Bayu Satriya Informatics Study Program, Faculty of Computer Science, Amikom University Yogyakarta Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta Telp: (0274) 884201 – 207 Email: riyanbayus1382@students.amikom.ac.id	

INTRODUCTION

The liver is a vital organ in the human body that performs a wide range of complex and essential functions. One of the diseases that affects the liver is hepatitis, commonly referred to as liver disease. According to the World Health Organization (WHO) Global Hepatitis Report 2017, nearly 1.2 million people, particularly in Southeast Asia and Africa, die each year due to liver-related illnesses (Nurlelah & Yuni Utami, 2022). Liver disease is a major global cause of morbidity and mortality. Various types of liver diseases exist, including fatty liver, cirrhosis, hepatitis, liver cancer, and chronic liver disease. One common cause of fatty liver is the excessive accumulation of triglyceride fat, which can lead to serious complications such as cirrhosis or liver cancer (Rekha Sundari et al., 2023).

In this study, liver disease is classified based on specific medical parameters provided in a research dataset. The dataset contains various patient medical records, such as liver enzyme levels (ALT, AST), bilirubin, albumin, total protein levels, and other risk factors related to liver health (Daely & Thomsen Nias, 2024). This data is used as input variables in the classification process to determine whether a person is affected by liver disease or not. Furthermore, the Random Forest algorithm is applied to classify patients based on the available medical parameters. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score to assess how effectively it can predict liver disease (Firdaus et al., 2024).

This research is grounded in the concept of machine learning, particularly in the fields of data mining and classification. Data mining is the process of extracting meaningful information from large datasets to uncover hidden patterns or relationships. One of the key techniques in data mining is classification, which is used to categorize data into specific classes based on patterns identified from historical data (Refindha et al., 2025). Random Forest is a tree-based classification method (Decision Tree), where an ensemble of decision trees is used to improve prediction accuracy. This algorithm offers several advantages, including robustness to missing values, the ability to handle data with numerous variables, and good performance in dealing with outliers (Herjanto & Carudin, 2024).

A study conducted by Kesuma et al. involved several steps, including data analysis, exploratory data analysis, preprocessing, algorithm modeling, and visualization. Based on these research stages, the prediction results for liver disease using the Random Forest algorithm showed an accuracy score of 0.713326 with an F1-score of 81% (Kesuma et al., 2023).

Another study by Lia et al. reported that the Random Forest algorithm achieved a classification accuracy of 78.63% in predicting liver disease. It was concluded that this accuracy outperformed other algorithms used in the same classification task (Lia et al., 2023).

A study conducted by Widya Kayohana utilized secondary data obtained from the internet, consisting of 15 variables including gender, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose, and one class label called Heart_stroke. The study compared several classification methods: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Stochastic Gradient Descent (SGD), which achieved accuracy scores of 68%, 82%, 90%, and 48%, respectively. Based on the comparison of accuracy levels, the Random Forest method achieved the highest accuracy, while Stochastic Gradient Descent recorded the lowest (Widya Kayohana, 2024).

This study aims to implement the Random Forest algorithm, optimized using Random Search, to predict liver disease and enhance efficiency and accuracy in disease diagnosis (Paunović-Pantić et al., 2024). The dataset used in this research is sourced from Kaggle, containing medical information related to patients suspected of having liver disease. The results of this study are expected to improve the accuracy of early liver disease detection, thereby contributing to medical decision support systems. Through a machine learning-based approach, the developed model can serve as a valuable tool in medical analysis, enhancing the speed and precision of liver disease detection (Nugroho et al., 2024).

\

RESEARCH METHODS

In this study, Random Search is used to optimize the Random Forest method in predicting liver disease. The sequence of this research flow is illustrated in the following figure.

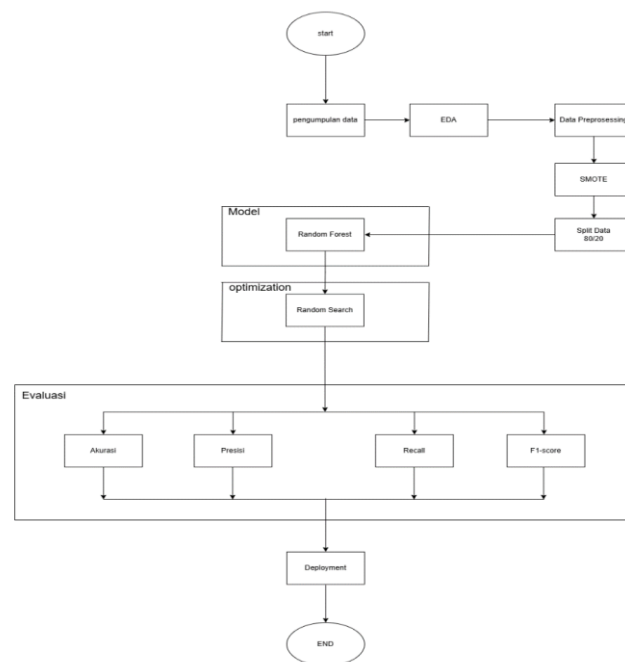


Figure 1. Research Flow

This research flow is explained as follows:

Data Collection

The dataset used in this study is a secondary dataset titled “Predict Liver Disease: 1700 Records Dataset”, obtained from the Kaggle Open Datasets and Machine Learning Project platform (Kharoua, 2024). The dataset consists of 1,700 rows and 11 features. It includes information on demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and liver disease diagnoses.

Exploratory Data Analyst (EDA)

The function of Exploratory Data Analysis (EDA) in liver disease prediction is to understand patterns, detect anomalies, examine data distribution, and identify relationships between variables in order to prepare the data before building an accurate predictive model.

Data Preprocessing

The data preprocessing stage is the process of preparing the dataset before it is used to train the model. Its purpose is to determine which columns will be used as variable X (features). Once these columns are identified, feature selection is performed to choose the features that have the strongest correlation with the target, so they can be used as inputs (X) in the liver disease prediction model.

SMOTE

The next step after data splitting is handling imbalanced data. This step is crucial to ensure the data is more balanced, allowing the data splitting process to run optimally and resulting in a more accurate model.

Split Data

After SMOTE is completed, the dataset is split into training data (80%) and testing data (20%) using the `train_test_split()` function from the `sklearn` library. This split ensures that the model can learn from the majority of the data, while the remaining 20% is used to evaluate the model's performance on previously unseen data. The 80% training portion

allows the model to recognize patterns across various patient data, and evaluation on the testing data provides a more objective measure of the model's predictive accuracy.

Modelling

To randomly search for the best combination of hyperparameters for the Random Forest algorithm, so that the resulting model achieves better performance in predicting liver disease. Random Search speeds up the tuning process compared to Grid Search because it does not test all possible combinations, yet it remains effective in finding parameters that yield high accuracy.

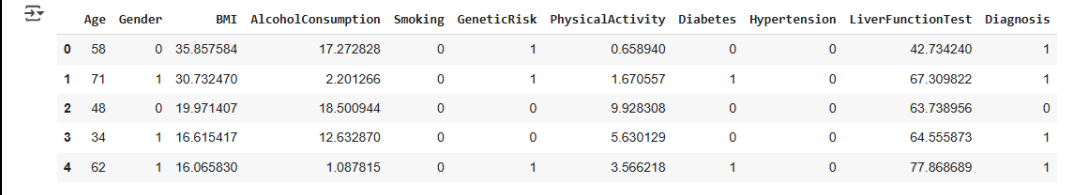
Evaluation

Evaluation is the stage used to assess the performance of the developed model. One of the evaluation methods that can be used is the confusion matrix. A confusion matrix is a method used to evaluate the performance of a classification model in machine learning, allowing for the calculation of how well the model correctly predicts the target classes. This table helps identify the types of errors made by the model and provides insights into where the model performs well or poorly. A confusion matrix typically consists of four components: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

RESULTS AND DISCUSSION

Data Collection

In the initial stage of this research, the dataset was obtained from Kaggle, an open platform that provides a wide variety of datasets. This study utilized a dataset titled *"Predict Liver Disease: 1700 Records Dataset"*, which was sourced from the Kaggle Open Datasets and Machine Learning Project. The dataset consists of 1,700 rows and 11 features. It includes demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and liver disease diagnoses.



	Age	Gender	BMI	AlcoholConsumption	Smoking	GeneticRisk	PhysicalActivity	Diabetes	Hypertension	LiverFunctionTest	Diagnosis
0	58	0	35.857584	17.272828	0	1	0.658940	0	0	42.734240	1
1	71	1	30.732470	2.201266	0	1	1.670557	1	0	67.309822	1
2	48	0	19.971407	18.500944	0	0	9.928308	0	0	63.738956	0
3	34	1	16.615417	12.632870	0	0	5.630129	0	0	64.555873	1
4	62	1	16.065830	1.087815	0	1	3.566218	1	0	77.868689	1

Figure 2. Data Import

Figure 2 displays the first five rows of the dataset using the `dataset.head()` function, showing that each column contains values consistent with the feature descriptions. During the data preprocessing stage, a correlation analysis will be performed to determine which features are suitable to be used as input variables (X). The target variable (y) has been defined in advance and is located in the Diagnosis column.

EDA.

In the Exploratory Data Analysis (EDA) stage, an initial analysis was conducted to understand the structure, distribution, and quality of the dataset. The dataset structure was examined using the `dataset.info()` function, which revealed that the dataset consists of 1,700 entries and a total of 11 columns. The data types include 4 columns of type float64 and 7 columns of type int64. No missing (null) values were found in the dataset, indicating that all columns contain complete data and are ready for further processing.

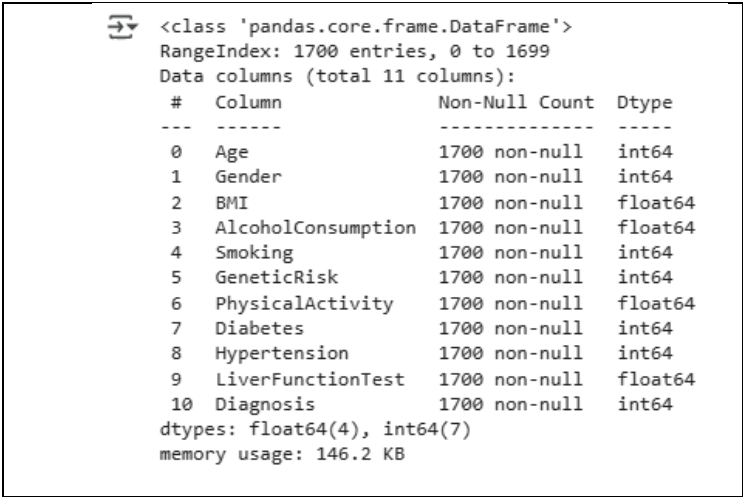


Figure 3. Dataset Structure Information

Figure 3 presents a summary of the dataset structure, including the number of rows and columns, the count of non-null values, and the data types (int64, float64, etc.).

The descriptive statistics of the dataset, obtained using dataset.describe(), provide an overview of the mean, minimum, maximum, and standard deviation values of the numerical columns.

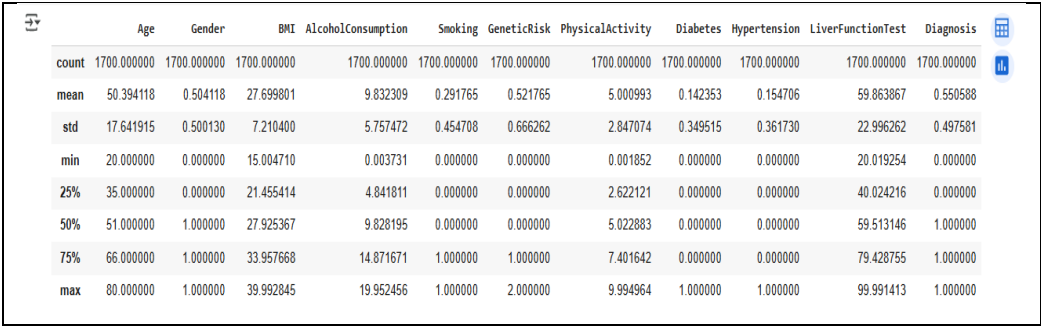


Figure 4. Descriptive Statistics of the Dataset

Figure 4 displays the descriptive statistics of the dataset, including the count, mean, standard deviation (std), minimum, maximum, and percentiles (25%, 50%, 75%) for each numerical feature in the dataset.

Data Preprocessing

In the data preprocessing stage, irrelevant columns were removed and feature selection was performed based on correlation analysis. To determine whether a correlation is strong or not, several steps must be carried out, including:

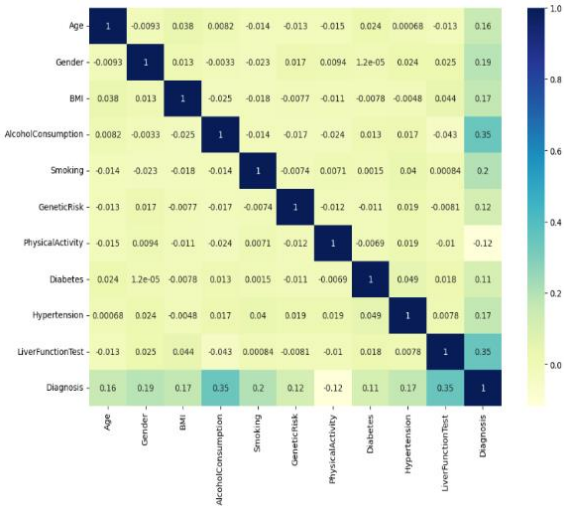


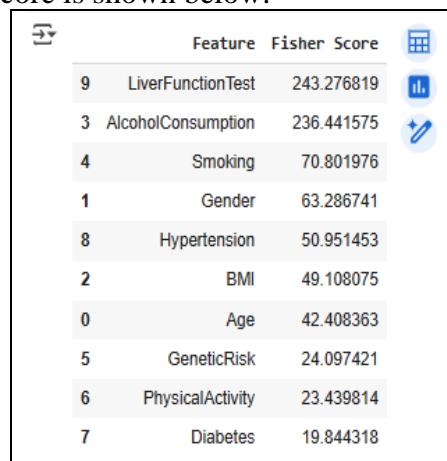
Figure 5. Heatmap of Feature Correlation Matrix

Figure 5 presents a heatmap of the correlation matrix between features, illustrating the statistical relationships among various risk factors and liver disease conditions. It can be observed that alcohol consumption (Alcohol Consumption) and liver function test results (Liver Function Test) show the strongest correlation with liver disease diagnosis, each with a correlation coefficient of 0.35. This indicates that these two variables are important predictors in determining liver health status. Other factors such as smoking habits (Smoking) also show a relatively significant correlation of 0.2, while patient age (Age) contributes moderately with a correlation value of 0.16.

Interestingly, this analysis also reveals that some factors, such as gender (Gender) and physical activity (Physical Activity), have relatively little influence on the diagnosis, with correlation values below 0.02. Although there are some relationships between independent variables, such as the correlation of 0.049 between hypertension and diabetes, no strong multicollinearity was found among the features overall. These findings provide valuable guidance in selecting variables for the prediction model, where alcohol consumption and liver function tests should be the primary focus, while some other variables may be considered for reduction in weight during further analysis.

Feature Selection

If no strong correlation is found, the next step is Feature Selection. In this study, the feature selection method used is Fisher Score. Fisher Score selects medical features that show significant differences between liver and non-liver patients. The first step in calculating the Fisher Score is shown below:

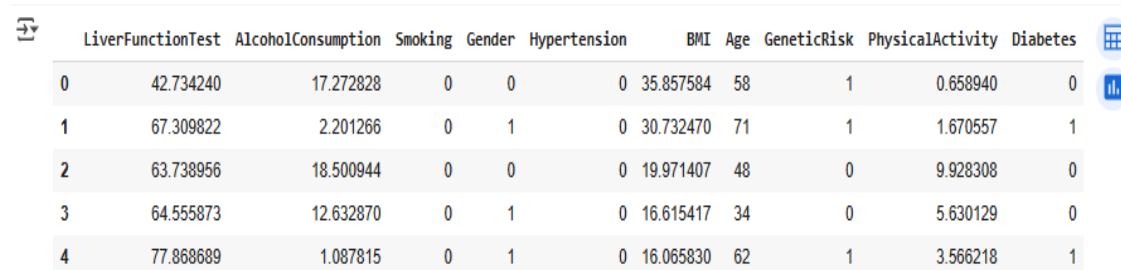


	Feature	Fisher Score
9	LiverFunctionTest	243.276819
3	AlcoholConsumption	236.441575
4	Smoking	70.801976
1	Gender	63.286741
8	Hypertension	50.951453
2	BMI	49.108075
0	Age	42.408363
5	GeneticRisk	24.097421
6	PhysicalActivity	23.439814
7	Diabetes	19.844318

Figure 6. Fisher Score Calculation Results

Figure 6 displays the results of feature selection based on Fisher Score, which helps in determining the most influential features for liver disease prediction.

After calculating the Fisher Score, the researcher needs to select a threshold to determine which features will be used in the modeling. In this study, the researcher chose a fixed threshold, where features with a Fisher Score > 1.0 are considered significant. Once the selection is made, the results are displayed using the command `X_selected.head()`, as shown below:



	LiverFunctionTest	AlcoholConsumption	Smoking	Gender	Hypertension	BMI	Age	GeneticRisk	PhysicalActivity	Diabetes
0	42.734240	17.272828	0	0	0	35.857584	58	1	0.658940	0
1	67.309822	2.201266	0	1	0	30.732470	71	1	1.670557	1
2	63.738956	18.500944	0	0	0	19.971407	48	0	9.928308	0
3	64.555873	12.632870	0	1	0	16.615417	34	0	5.630129	0
4	77.868689	1.087815	0	1	0	16.065830	62	1	3.566218	1

Figure 7. Modeling Based on Fixed Threshold (Feature Selection Results)

Figure 7 displays the results obtained after performing Feature Selection using the Fisher Score method, highlighting the best features selected, including: 'Age', 'Gender', 'BMI', 'AlcoholConsumption', 'Smoking', 'GeneticRisk', 'PhysicalActivity', 'Diabetes', 'Hypertension', and 'LiverFunctionTest'.

SMOTE

At this stage, handling of imbalanced data is performed using the Synthetic Minority Oversampling Technique (SMOTE). Data imbalance occurs when the number of samples in one target class is much smaller compared to the other class.

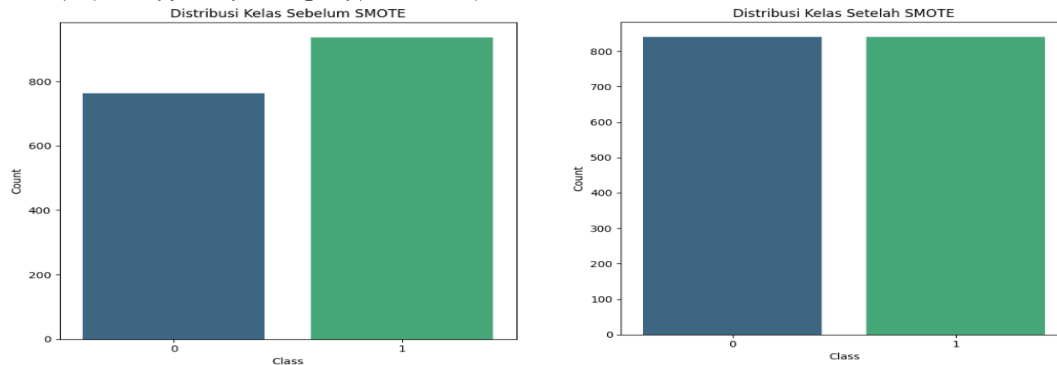


Figure 8. Class Distribution Left Before SMOTE Right After SMOTE

Figure 8 (left) shows the class distribution analysis before applying SMOTE, where it can be observed that the "Liver" class has a significantly larger number of samples compared to the "NonLiver" class. This imbalance can lead the machine learning model to be biased towards predicting the majority class, thereby reducing the model's ability to accurately recognize the target class for Diagnosis.

After applying SMOTE, as shown in Figure 8 (right), the class distribution becomes balanced. SMOTE works by creating synthetic data for the Diagnosis class by averaging the nearest neighbors to generate new samples. This increases the number of samples in the "NonLiver" class to match the "Liver" class. With a more balanced data distribution, the model is expected to better learn both classes, thereby improving prediction performance, particularly for the Diagnosis class.

Split Data

At this stage, the dataset is split into training data (80%) and testing data (20%) using the `train_test_split` function from the `sklearn` library. The split is performed while maintaining the proportion of the target class to ensure that the distribution between the training and testing data remains representative.

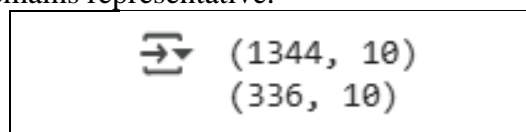


Figure 9. Results of X_train and X_test

Figure 9 shows the results of this process, which yields a total of 1,344 samples for the training data and 336 samples for the testing data for both variables (X and y). The results of the data split are displayed in the following Table 1.

Table 1. Number of Training and Testing Data Samples

Data	Number of Samples.
Data training (X_train)	1344
Data testing (X_test)	336
Data training (y_train)	1344
Data testing (y_test)	336

Random Forest Model (Without Optimization)

At this stage, a Random Forest model is created and evaluated without hyperparameter optimization. The model is trained using the preprocessed training data and tested on the testing data to measure its prediction performance.

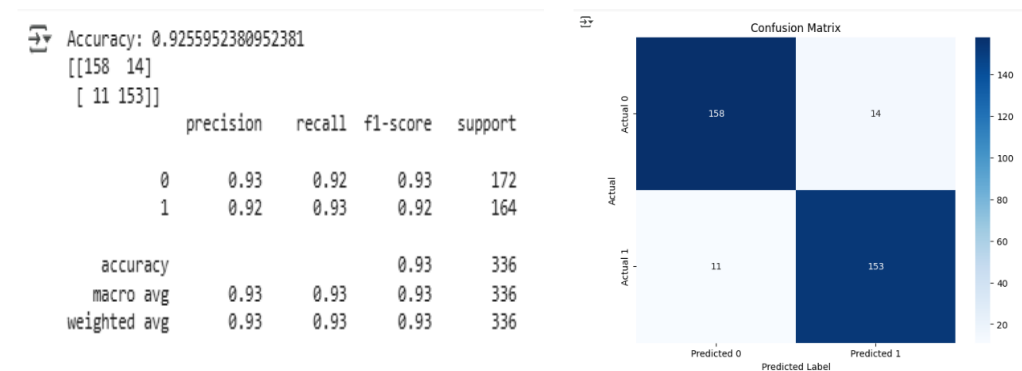


Figure 10. Random Forest Model Evaluation and Confusion Matrix Without Optimization

Based on the evaluation results displayed in Figure 10, the model shows good performance in detecting liver disease. The evaluation results indicate that the model achieves an accuracy of 93%, with precision, recall, and F1-score values of 92%, 93%, and 92%, respectively, for the "Liver" class. For the "Non-Liver" class, the model has a precision of 93%, recall of 92%, and F1-score of 93%. The high values for these metrics demonstrate that the model handles both classes well.

Random Forest Model Optimization Using Random Search

After evaluating the Random Forest model without optimization, the next step is to perform hyperparameter optimization using the Random Search method. This optimization aims to find the best combination of hyperparameters that can improve the model's performance in detecting liver disease.

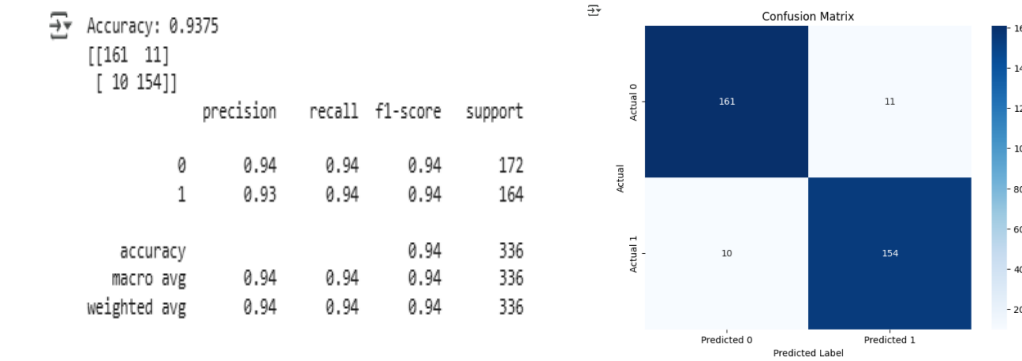


Figure 11. Random Forest Model Evaluation and Confusion Matrix Using Random Search Optimization

The evaluation matrix in Figure 11 shows an improvement in performance compared to the previous model. The model achieves an accuracy of 94%, with precision, recall, and F1-score values of 93%, 94%, and 94%, respectively, for the "Liver" class. Meanwhile, for the "NonLiver" class, the precision and recall values are 94% and 94%, with an F1-score of 94%. The average metrics also show consistent results with high accuracy across all classes.

Evaluation

At this stage, an analysis and comparison of the evaluation results of the Random Forest model before and after hyperparameter optimization using Random Search are performed. The evaluation includes metrics such as accuracy, precision, recall, and F1-score to measure the model's performance in detecting liver disease. The evaluation results are summarized in Table 2 below.

Table 2. Comparison of Model Evaluation Results Before and After Optimization

Metrics	Before Optimization	After Optimization	Ascension
Accuracy	93%	94%	+1%
Precision	92%	93%	+1%
Recall	93%	94%	+1%
F1-Score	92%	94%	+2%

From Table 2 above, it can be seen that hyperparameter optimization using Random Search resulted in an improvement in the model's performance. The accuracy increased from 93% to 94%. Precision improved to 93%, recall remained at 94%, and the F1-score increased to 94%.

This evaluation shows that hyperparameter optimization using Random Search successfully improved the overall model performance, particularly in reducing the number of prediction errors for both classes. Therefore, the optimized model can be used to detect liver disease with higher accuracy and reliability.

Deployment Results

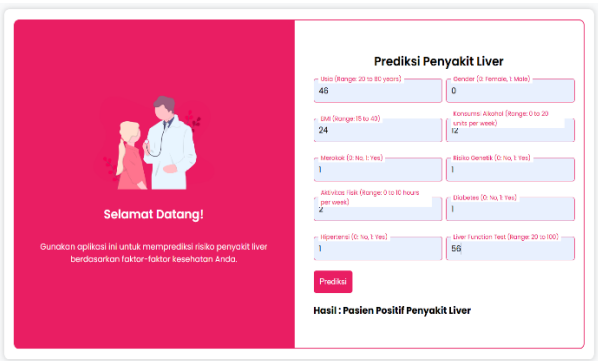


Figure 12. Deployment Results

Figure 12 shows that this liver disease prediction application provides users with an easy way to assess the risk of liver disorders based on their health profile independently. By entering data such as age, gender, body mass index (BMI), alcohol consumption habits, physical activity levels, as well as a history of diabetes and hypertension, the system analyzes the likelihood of liver issues. In the example results displayed, the application predicts "Positive Liver Disease Patient," indicating that the combination of input parameters—such as high BMI, significant alcohol consumption, or abnormal liver function test results—suggests a potential risk of liver disorders that should be monitored.

The prediction results should be viewed as an early warning, encouraging users to seek further consultation with a professional healthcare provider. This application is designed as an initial screening tool, not as a definitive diagnostic tool. To improve prediction accuracy, it is recommended to supplement the data with the latest laboratory test results and additional health information such as family history. In this way, this tool can serve as a useful first step in the early detection of liver issues before undergoing more comprehensive medical examinations.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results of the research conducted, it can be concluded that hyperparameter optimization using the Random Search method successfully improved the performance of the Random Forest algorithm in detecting liver disease. The Random Forest model without optimization initially showed good performance with an accuracy of 93%, as well as high precision, recall, and F1-score values for both classes, especially after applying the SMOTE method to address data imbalance.

After hyperparameter optimization using Random Search, the model's performance improved with accuracy rising to 94%, as well as an increase in precision and F1-score, particularly for the "Liver" class, which reached 93% and 94%, respectively. This improvement demonstrates that selecting optimal hyperparameters can enhance the model's ability to better recognize data patterns, especially in detecting the Diagnosis class (Liver).

Thus, this research proves that hyperparameter optimization using Random Search can significantly improve the performance of Random Forest in early prediction of liver disease. The results of this study are expected to serve as a reference in the development of machine learning-based diagnostic methods to assist healthcare professionals in making faster and more accurate liver disease predictions.

Suggestion

Based on the results of the research conducted, here are some recommendations that can serve as a guide for further development and improvements in the future:

In addition to Random Forest, other algorithms such as XGBoost, Gradient Boosting, or Deep Learning could be explored to compare better performance in detecting liver diseases. This could provide additional insights into which algorithm is more effective and efficient for this type of data.

Hyperparameter optimization in this study was performed using Random Search. Future research could explore other approaches such as Grid Search or Bayesian Optimization to explore hyperparameter combinations more thoroughly and efficiently.

For future research, model interpretability analysis can be conducted using techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) to understand which features have the most influence on the model's predictions.

REFERENCE

- Daely, P. J., & Thomsen Nias, R. D. M. (2024). Laporan Kasus: Pemberian Hepatoprotektor Dan Anti Oksidan Pada Perlemakan Hati Non Alkaholik (NAFLD).
- Firdaus, R., Habibie, H., Rizki, Y., Informatika, T., Komputer, I., Riau, U. M., & Id, R. A. (2024). Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data.
- Herjanto, M. F. Y., & Carudin, C. (2024). ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI SIREKAP PADA PLAY STORE MENGGUNAKAN ALGORITMA RANDOM FOREST CLASSIFER. *Jurnal Informatika Dan Teknik Elektro Terapan*, 12(2). <https://doi.org/10.23960/jitet.v12i2.4192>
- Kesuma, M., Informatika dan Bisnis Darmajaya, I., Pagar Alam No, J. Z., & Lampung, B. (2023). PREDIKSI PENYAKIT LIVER MENGGUNAKAN ALGORITMA RANDOM FOREST. *Jurnal Informasi Dan Komputer*, 11(2).
- Kharoua, R. El. (2024). Predict Liver Disease: 1700 Records Dataset. <https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset/data>
- Lia, F., Cahyanti, D., Sarasati, F., Astuti, W., & Firasari, E. (2023). KLASIFIKASI DATA MINING DENGAN ALGORITMA MACHINE LARNING UNTUK PREDIKSI PENYAKIT LIVER. In *Technologia* (Vol. 14, Issue 2). <https://ojs.uniska-bjm.ac.id/index.php/JIT>

- Nugroho, R. E., Pamungkas, W. Y., & Jaman, J. H. (2024). PENDETEKSI PENYAKIT HEPATITIS MENGGUNAKAN CART DECISION TREE. *Jurnal Informatika Dan Teknik Elektro Terapan*, 12(3S1). <https://doi.org/10.23960/jitet.v12i3S1.5184>
- Nurlelah, E., & Yuni Utami, D. (2022). SELEKSI ATRIBUT PADA ALGORITMA NEURAL NETWORK MENGGUNAKAN PARTICLE SWARM OPTIMIZATION UNTUK DIAGNOSIS PENYAKIT LIVER. In *Daerah Khusus Ibukota Jakarta* (Vol. 98, Issue 9).
- Paunović-Pantić, J., Vučević, D., Pantić, I., Valjarević, S., & Radosavljević, T. (2024). Development of random forest machine learning model for the detection of changes in liver tissue after exposure to iron oxide nanoparticles. *Medicinska Istrazivanja*, 57(1), 21–26. <https://doi.org/10.5937/medi57-46969>
- Refindha, F., Harianto, A., Alawi, Z., & Sa'ida, I. A. (2025). PENGARUH KOMPOSISI SPLIT DATA PADA AKURASI KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN ALGORITMA MACHINE LEARNING. *Jurnal Sistem Informasi Dan Informatika (Simika)*, 8(1).
- Rekha Sundari, M., Raga Manognya, P., Venkata Mahesh, A., & Swetha, M. (2023). PREDICTION OF LIVER DISEASES WITH RANDOM FOREST CLASSIFIER WITH PRINCIPAL COMPONENT FEATURE EXTRACTION. 18(17). www.arpnjournals.com
- Widya Kayohana, K. (2024). KLASIFIKASI PENYAKIT HATI MENGGUNAKAN RANDOM FOREST DAN KNN. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 4).