Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1159~1164
e-ISSN: 2548-6861

Random Forest-based Hepatocellular Carcinoma Liver Disease Classification Model with LDA Feature Selection on Patient Medical Records

Nurul Istiqamah [1]*, Arif Iman Anshori [2], Novita Rahmayuna [3]***, Umi Meganinditya Wulandari [4]****
* Information Systems and Technology, Nobel Institute of Technology and Business Indonesia
** Informatics, Nahdlatul Ulama Institute of Technology and Science, Pekalongan
*** Information Systems, Bina Nusantara University
nrlistiqamah127@gmail.com [1], arifaan82@gmail.com [2], novita.rahmayuna@binus.ac.id [3], nindityaw@gmail.com[4]

Article Info

Article history:

Received 2025-10-24
Revised 2026-02-17
Accepted 2026-02-27

Keyword:

Hepatocellular Carcinoma
Random forest
Feature selection
Classification
LDA.

ABSTRACT

Hepatocellular carcinoma (HCC) is one of the leading causes of liver cancer mortality worldwide, and early detection remains challenging due to the complexity of clinical indicators. This study investigates a Random Forest-based classification model for HCC using patient medical record data, with Linear Discriminant Analysis (LDA) applied as a feature selection approach. The dataset consists of 100 clinical records comprising 39 attributes. A stratified 80:20 train–test split and cross-validation were employed to evaluate model stability. The baseline Random Forest model achieved an accuracy of 85% with an AUC of 0.69, indicating moderate discrimination performance. When LDA-based feature selection was applied prior to classification, predictive performance did not improve under the current dataset conditions. Although LDA contributed to identifying clinically relevant variables such as bilirubin markers and viral infection indicators, dimensionality reduction did not enhance overall classification results. These findings suggest that Random Forest provides relatively stable performance for HCC classification within limited datasets, while LDA-based feature selection primarily contributes to interpretability rather than predictive gain. However, the results should be interpreted cautiously due to the small sample size and class imbalance. Future work should involve larger datasets and rigorous validation strategies to improve generalization capability.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Feature Selection (FS) is a data pre-processing technique aimed at reducing the number of features or variables in a dataset by selecting only those that are relevant and informative to the target prediction. This process involves eliminating features with low predictive power, as well as those that are irrelevant or redundant [1]. The exponential growth in data acquisition and storage has led to datasets with an extensive number of features. While, in theory, an increased number of features can enhance a model's discriminative capabilities, in practice, it often introduces several challenges. These include heightened complexity, increased computational cost, and the presence of irrelevant or redundant features. Such conditions may degrade the model's accuracy in classifying new data, as the estimation of classification errors becomes suboptimal.

To address these challenges, feature selection serves as a crucial step by filtering only the most relevant and informative features, thereby producing a more efficient and accurate classification model [2]. FS techniques are generally classified into three categories: (1) Filter methods; (2) Wrapper methods; and (3) Embedded methods [3]. The application of FS has been widely demonstrated across various domains, including Intrusion Detection Systems (IDS)[4], breast cancer diagnosis [5], credit card fraud detection [6], and text classification [7]. The rationale for applying FS is that classification models tend to perform better and produce more consistent outcomes when trained on a smaller subset of informative features, rather than on all

available features [7]. FS has thus become a vital component in the machine learning workflow [8]. In many high-dimensional datasets, only a fraction of the available features are actually relevant to data mining and machine learning tasks [9].

Hepatocellular carcinoma (HCC) is one of the most prevalent and life-threatening forms of liver cancer worldwide, accounting for approximately 90% of primary liver cancer cases. Early diagnosis of HCC remains a significant clinical challenge due to the disease's asymptomatic progression and the complexity of biochemical indicators associated with liver dysfunction [3]. In this context, machine learning models have shown promising potential for improving diagnostic accuracy by identifying subtle patterns in clinical and laboratory data. This study aims to evaluate the effectiveness of Linear Discriminant Analysis (LDA) in the feature selection process, specifically in identifying the most influential features in Hepatocellular Carcinoma (HCC) disease data. The methodological approach in this study involves two primary stages. While LDA is not a pure feature selection method like filter or wrapper techniques, it is more accurately characterized as a feature extraction or dimensionality reduction technique. Nevertheless, LDA can still be utilized to identify features that are highly discriminative for class separation. The feature selection process using LDA consists of two main phases: the discriminant analysis phase and the dimensionality reduction phase. Subsequently, the selected features are used in the classification stage, which employs the Random Forest (RF) algorithm.

II. METHOD

A. Data Preparation

This study utilizes data on Hepatocellular Carcinoma (HCC) of the liver, consisting of 100 records obtained from medical documentation. The dataset is categorized as heterogeneous, as it reflects clinical variability among individuals. Each entry comprises 39 independent attributes representing the patients' medical and demographic features. The dependent variable is a binary survival label, divided into two classes: class 0 indicates patients who did not survive, while class 1 represents patients who successfully survived.

Missing data were handled using mode imputation for categorical variables and mean imputation for numerical variables. This strategy was applied to maintain the overall characteristics of the dataset while minimizing potential distortion. Binary indicators, including HBsAg and HCVAb, were encoded as 0 (negative) and 1 (positive). In addition, feature standardization was performed to ensure that variables measured on different scales contributed proportionally during the modelling process.

To obtain a balanced evaluation, the dataset was divided into training and testing subsets using a stratified 80:20 split. This approach preserves the original class distribution in both subsets and helps reduce potential bias in performance estimation, particularly given the relatively limited size of the dataset.

The resulting dataset thus provides a reliable foundation for the subsequent stages of feature selection and classification model development.

B. Feature Selection with Linear Discriminant Analysis

Feature selection (FS) is an important step in reducing data dimensionality by identifying a subset of relevant features from the original dataset. This procedure is conducted during the preprocessing stage to enhance This procedure is conducted. In general, there are three main categories of FS approaches. The first is the filter method, which employs a "proxy measure" derived from the general characteristics of the training data to evaluate individual features or feature subsets prior to modeling. The second is the wrapper method, an iterative and computationally intensive process capable of identifying the optimal feature set for a specific modeling algorithm. The third is the embedded method, which integrates feature selection directly within the model training process [3], [10].

One of the techniques used to simplify classification problems in supervised machine learning is Linear Discriminant Analysis (LDA). This technique is used to model the differences between two or more groups, as it requires distinguishing among multiple classes. It enables the transformation of features from a high-dimensional space into a lower-dimensional space [11], [12].

In this study, Linear Discriminant Analysis (LDA) was implemented as an exploratory feature selection approach. LDA transforms features into a lower-dimensional space by minimizing the ratio of between-class variance to within-class variance. The model was fitted solely on the training dataset, and the resulting transformation was then applied to the testing dataset to avoid data leakage.

C. Classification with Random Forest

Random Forest is a supervised machine learning approach developed as an extension of the basic classification algorithm, Decision Tree. In this tree-based ensemble learning method, each tree is constructed using a randomly selected subset of variables. It combines two distinct tree-based strategies in two critical aspects. Each tree is first trained on a randomly selected bootstrap sample drawn from the entire dataset [13].

In ensemble methods such as Random Forest, final predictions are produced through a majority voting mechanism across multiple decision trees [14]. In the context of ensemble algorithms such as Random Forest, this voting mechanism serves as an aggregation strategy that effectively reduces variance and minimizes the risk of errors that may occur when relying on a single decision tree. Each decision tree is constructed from a different subset of data through a bootstrap sampling process, thereby providing diverse perspectives on the underlying data distribution.

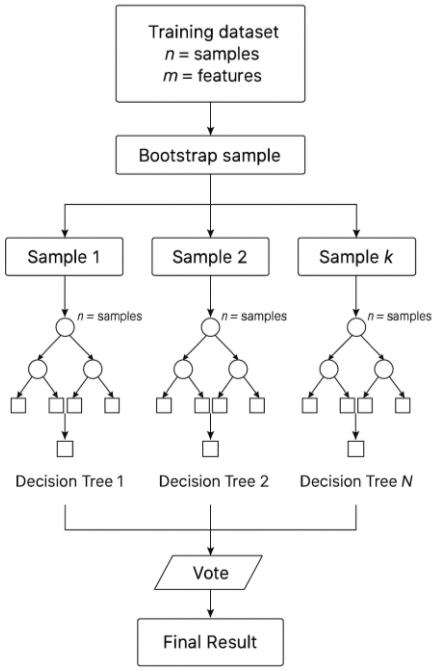

Figure 1. Random Forest Classifier Architecture [13]

In this study, Random Forest was selected due to its ability to handle heterogeneous clinical variables and its relatively stable performance when applied to limited medical datasets.

D. Confusion Matrix

A confusion matrix is a tool used to evaluate model performance in machine learning, particularly for classification tasks [15]. This matrix compares the model's predicted outcomes with the test data and presents the corresponding true values.

TABLE 1.
CONFUSION MATRIX

	Predicted (-)	**Predicted (+)**
Actual (-)	TN	FP
Actual (+)	FN	TP

The abbreviations TP, TN, FP, and FN in Figure 2 represent the outcomes of each class in the classification procedure. True Positive (TP) indicates the condition in which the model correctly predicts a data instance as positive, matching the actual positive condition. True Negative (TN) occurs when the model classifies a data instance as negative, and this prediction aligns with the actual negative condition. Conversely, False Positive (FP) describes a situation where the model predicts a data instance as positive when it is actually negative; this condition is commonly associated with a Type I Error. Meanwhile, False Negative (FN) refers to a condition in which the model predicts a data instance as negative, but in reality, it belongs to the positive class; this is known as a Type II Error [14]. The calculations for accuracy, precision, and recall can be determined using the formulas presented in Equations (1), (2), and (3).

$$\text{Accuracy} = (TP + TN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

III. RESULTS AND DISCUSSION

A. Workflow

The workflow of this study, as illustrated in Figure 3, consists of three main stages: data preprocessing, data preparation, and data classification.

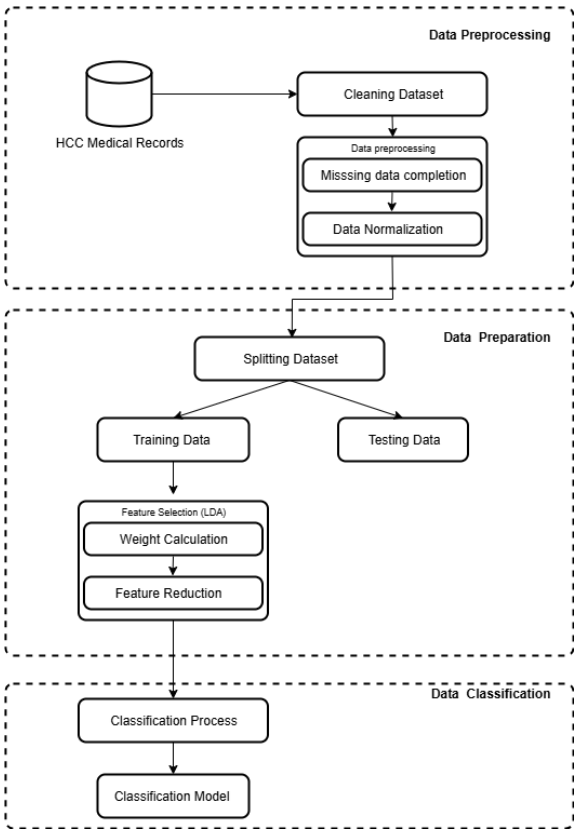

Figure 3. Explanation Research Workflow RF+LDA

The process begins with the data preprocessing stage, which aims to transform the raw medical records of Hepatocellular Carcinoma (HCC) into a clean and structured dataset suitable for analysis. The first step in this stage is data cleaning, which serves to eliminate inconsistent, redundant, or irrelevant data. This is followed by missing data completion, which addresses incomplete attribute values to ensure that no critical information is overlooked. Subsequently, data normalization is performed to standardize the value ranges across variables, thereby preventing bias caused by differences in scale.

After the preprocessing stage, the dataset was partitioned into training and testing subsets using a stratified 80:20 split to preserve the original class distribution. Feature selection using Linear Discriminant Analysis (LDA) was then performed on the training data to identify discriminative

attributes with respect to the target class. The learned transformation was subsequently applied to the testing data to maintain methodological consistency and prevent data leakage.

Following the feature selection process, the training data were used to construct the model, while the testing data were employed to evaluate its performance and generalization capability. The final stage involved model construction and evaluation. A Random Forest classifier was trained using the LDA-selected features derived from the training data and evaluated on the testing data to assess predictive performance.

B. Implementation Method

Prior to the model construction phase, as presented in Table 1, approximately 39 features were processed using Linear Discriminant Analysis (LDA). The application of LDA aimed to identify the most effective linear combinations of variables for distinguishing between the two target classes.

The LDA model was applied to the normalized training data to extract a principal discriminative component. The resulting transformation coefficients were used to rank features based on their contribution to class separation. Features with higher absolute coefficient values were considered more influential in differentiating between the two classes.

Based on this ranking process, the original feature set was reduced to the top ten features with the highest absolute coefficient values. In this study, LDA is employed as an exploratory feature selection approach rather than as a strict statistical optimization method. By reducing the feature space, the model complexity was lowered while retaining clinically relevant discriminatory information for the subsequent classification stage.

C. Result

This study divided the feature-selected dataset into 80% training data and 20% testing data. The feature reduction process using Feature Selection (FS) with LDA produced the ten best-ranked. The selected top ten features are presented in Table 2. Feature ranking was determined based on absolute LDA coefficients.

TABLE 2.
10 BEST FEATURES

Rank	Features	Explanation
21	BilD	Direct bilirubin
22	BilT	Total bilirubin
16	MCV	The hemoglobin level in each red blood cell
34	Asci	Ascites degree
31	HCVAb	HCV (Hepatitis C Virus Antibody) test results
20	SGPT	Alanine transaminase
32	Spleno	Splenomegaly
29	HBsAg	HBeAg (Hepatitis B e Antigen) test results
15	MCH	Red blood cell count
25	Trom	Platelets

Two experimental settings were examined: (1) Random Forest and (2) Random Forest combined with LDA-based feature selection. This study employed the Random Forest Classifier algorithm as the primary model to distinguish between Hepatocellular Carcinoma (HCC) and non-HCC patients. This ensemble learning method combines multiple decision trees to improve prediction accuracy and stability. Each tree is constructed from a randomly selected subset of data and features using the bootstrap sampling technique. The model was trained with key parameters set to n_estimators = 100 and random_state = 42. The combination of optimal feature selection using LDA and training with the Random Forest algorithm produced a model whose evaluation results are presented in Table 3 and Table 4:

TABLE 3.
ACCURACY OF MODELS

Model	Accuracy
RF	85%
LDA + RF	55%

TABLE 4.
EVALUATION PARAMETERS MODEL

Model	Precision	Recall
RF	89%	94%
LDA + RF	71%	67%

TABLE 5.
EVALUATION CROSS-VALIDATION MODEL

Model	AUC
RF	0.69
LDA + RF	0.68

Precision, recall, and accuracy metrics were used to evaluate the model's performance. As shown in Tables 3 and 4, the Random Forest model using the full feature set achieved an accuracy of 85% on the testing dataset. The precision and recall for the positive class were 0.89 and 0.94, respectively. These results indicate that most positive cases in the testing subset were correctly classified. Additionally, the Random Forest model combined with LDA-based feature selection achieved an accuracy of 55%, with precision and recall values of 0.71 and 0.67, respectively. This decrease in performance suggests that dimensionality reduction using LDA did not improve classification results in this dataset.

As presented in Table 5, the Curve (AUC) values for both models. The Random Forest model obtained an AUC of 0.69, indicating moderate class discrimination. The LDA-enhanced model obtained an AUC of 0.68 in cross-validation, showing comparable but variable discrimination performance.

Figure 4 presents the confusion matrix of the baseline Random Forest model. The model successfully classified 16 positive cases (True Positive) and 1 negative case (True Negative) correctly, while 1 positive case was misclassified as negative (False Negative) and 2 negative cases were incorrectly predicted as positive (False Positive). These results indicate that most positive cases in the test subset were

successfully identified with a relatively small number of classification errors.

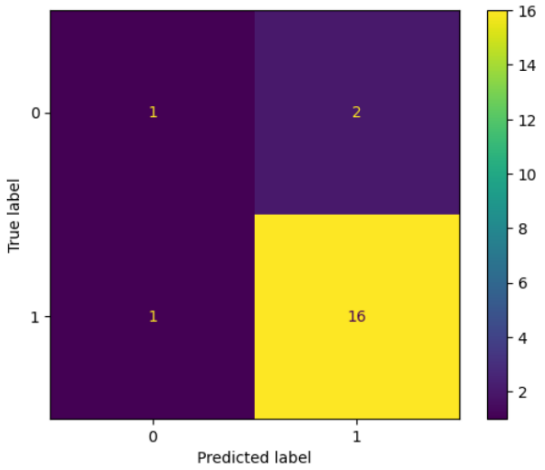

Figure 4. Evaluation Parameters using RF

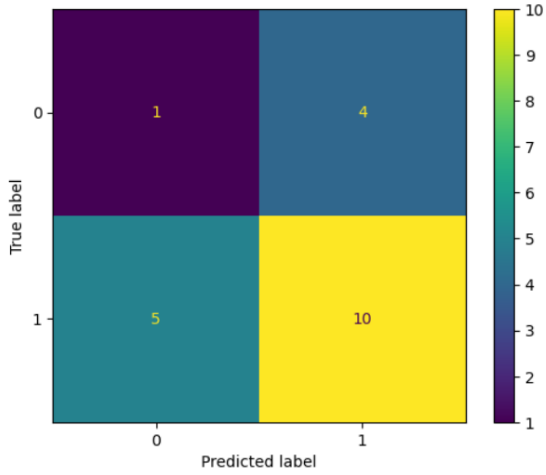

Figure 5. Evaluation Parameters using RF + LDA

Figure 5 illustrates the confusion matrix of the Random Forest model combined with LDA-based feature selection. In this model, 10 positive cases were correctly classified, while 5 positive cases were misclassified as negative. Additionally, 1 negative case was correctly classified and 4 negative cases were misclassified as positive. Compared to the baseline model, the LDA-enhanced model resulted in a higher number of false negatives and overall misclassifications.

Overall, the baseline Random Forest model demonstrated comparatively more stable performance than the LDA-enhanced configuration under the current experimental setting. However, this does not necessarily indicate that the baseline model represents an optimal or fully robust solution.

IV. CONCLUSIONS

This study integrates the Linear Discriminant Analysis (LDA) method for feature selection with the Random Forest Classifier algorithm as the classification model for Hepatocellular Carcinoma (HCC) liver disease. The experimental results indicate that, under the current dataset conditions, the baseline Random Forest model demonstrated relatively more stable performance compared to the configuration incorporating LDA-based feature reduction. Several key clinical variables identified through the LDA-based feature selection process include Direct Bilirubin (BilD), Total Bilirubin (BilT), SGPT, HCVAb, and HBsAg. Although LDA contributed to feature interpretability by highlighting discriminative attributes, it did not improve predictive performance within this dataset.

The baseline Random Forest model achieved moderate classification performance, as reflected by its accuracy and AUC values. However, these findings should be interpreted with caution due to the limited dataset and the presence of class imbalance. No additional resampling or class-weighting strategies were applied, which may influence model sensitivity and generalization capability

For future research development, it is recommended to expand the quantity and diversity of the dataset to enhance the model's generalization capability. Additionally, the application of advanced optimization techniques such as Principal Component Analysis (PCA) or Grid Search Hyperparameter Tuning can be employed to further improve predictive performance and reduce potential bias within the model.

REFERENCES

[1] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Front. Bioinforma.*, vol. 2, June 2022, doi: 10.3389/fbinf.2022.927312.

[2] Huan. Liu and Hiroshi. Motoda, *Computational methods of feature selection*. Chapman & Hall/CRC, 2008, p. 419.

[3] U. M. Wulandari, B. Warsito, and F. Farikin, "Survival Information System Using ReliefF Feature Selection and Backpropagation in Hepatocellular Carcinoma Disease," in *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, July 2023, pp. 37–42. doi: 10.1109/ISITIA59021.2023.10221079.

[4] Y. Yin *et al.*, "IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset," *J. Big Data*, vol. 10, no. 1, Feb. 2023, doi: 10.1186/s40537-023-00694-8.

[5] E. Odhiambo Omuya, G. Onyango Okeyo, and M. Waema Kimwele, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Syst. Appl.*, vol. 174, p. 114765, July 2021, doi: 10.1016/j.eswa.2021.114765.

[6] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00573-8.

[7] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.

[8] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Appl. Soft Comput.*, vol. 86, p. 105836, Jan. 2020, doi: 10.1016/j.asoc.2019.105836.

[9] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 1, Art. no. 1, May 2020, doi: 10.38094/jastt1224.

[10] H. H. Htun, M. Biehl, and N. Petkov, "Survey of feature selection and extraction techniques for stock market prediction," *Financ. Innov.*, vol. 9, no. 1, Jan. 2023, doi: 10.1186/s40854-022-00441-7.

[11] M. O. Adebiyi, M. O. Arowolo, M. D. Mshelia, and O. O. Olugbara, "A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis," *Appl. Sci.*, vol. 12, no. 22, Art. no. 22, Jan. 2022, doi: 10.3390/app122211455.

[12] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun*, vol. 30, no. 2, pp. 169–190, Jan. 2017, doi: 10.3233/AIC-170729.

[13] M. Park, D. Jung, S. Lee, and S. Park, "Heatwave Damage Prediction Using Random Forest Model in Korea," *Appl. Sci.*, vol. 10, no. 22, Art. no. 22, Jan. 2020, doi: 10.3390/app10228237.

[14] N. Istiqamah, B. Surarso, and B. Warsito, "Classification of customer review using random forest classifier," *AIP Conf. Proc.*, vol. 2738, no. 1, p. 060005, June 2023, doi: 10.1063/5.0140436.

[15] N. Rahmayuna, D. S. Rahardwika, C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "Pathogenic Bacteria Genus Classification using Support Vector Machine," in 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE, Nov. 2018, pp. 23–27. doi: 10.1109/ISRITI.2018.8864478