Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 1, January 2026, pp. 430~438
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i1.pp430-438

Comparative analysis of linear regression, random forest, and LightGBM for hepatitis disease prediction

Hennie Tuhuteru[1], Goldy Valendria Nivaan[1], Marvelous Marvel Rijoly[1], Joselina Tuhuteru[2]
[1]Faculty of Computer Science, Universitas Kristen Indonesia Maluku, Ambon, Indonesia
[2]Faculty of Economic and Business, Universitas Kristen Indonesia Maluku, Ambon, Indonesia

Article Info

Article history:

Received Jun 5, 2025
Revised Oct 6, 2025
Accepted Dec 22, 2025

Keywords:

Hepatitis
LightGBM
Linear regression
Random forest
Survival

ABSTRACT

In bioinformatics research, computational pattern-analysis techniques are frequently employed to assist in disease prediction and diagnostic modeling, including applications for hepatitis prognosis. Hepatitis is a type of serious disease with various types that have the potential to threaten the life of the sufferer without showing significant symptoms and signs, so many sufferers do not realize that they are affected by the disease. Various methods are used to predict diseases in the hope of providing the best results from the learning model used. The objective of this study is to implement linear regression, random forest, and light gradient boosting machine (LightGBM) to estimate mortality risk among hepatitis patients. In addition, a performance comparison of the results of hepatitis disease prediction using the three algorithms was also carried out to find out which model gave the most accurate and optimal results. The results of this study show that the application of learning models from the linear regression, random forest and Light-GBM algorithms has been successfully carried out to predict the survival status of patients with hepatitis. The findings reveal that random forest achieved the highest predictive performance with an accuracy of 84%, followed by LightGBM at 77% and linear regression at 32%.

Corresponding Author:

Hennie Tuhuteru
Faculty of Computer Science, Universitas Kristen Indonesia Maluku
Ambon, Indonesia
Email: hannytuhuteru@gmail.com

1. INTRODUCTION

The use of technology as a contemporary innovation to analyze and predict data patterns and trends has become an essential approach across various fields, particularly in medicine and global healthWithin this framework, data-driven analytical methods help process extensive clinical information collected through routine medical activities, allowing deeper exploration of disease-related trends. This approach integrates machine learning, advanced computing, and information retrieval, which have collectively transformed disease diagnosis and prediction in the field of bioinformatics [1]–[4].

One of the major diseases addressed through this technological application is hepatitis, a serious liver condition caused by viral infection, which often remains asymptomatic in its early stages. This silent progression leads to late diagnoses, posing a greater risk of complications such as cirrhosis and liver failure [5]–[8]. Common symptoms, including fever, nausea, fatigue, easy bruising, and jaundice, may not appear until advanced stages of liver damage [9], [10].

Recent reports from the world health organization (WHO), hepatitis caused an estimated 1.3 million deaths in 2022, with over 2.2 million new infections recorded in the same year. A total of 38 countries accounted for nearly 80% of global infections and deaths, with Indonesia ranked among the top 10 countries

with the highest hepatitis burden [11]. Predicting mortality and survival rates in hepatitis patients remains a significant challenge in efforts to improve the effectiveness of treatment and medical intervention. Predictive modeling not only supports early diagnostic insights but also substantially contributes to informed clinical decisions and the design of appropriate therapeutic strategies [12]. The incorporation of techniques such as synthetic minority over-sampling technique (SMOTE), support vector machine recursive feature elimination (SVM-RFE), and hyperparameter tuning has enhanced model performance in cases involving class imbalance or noisy data [13]–[15].

Studies have repeatedly shown that the random forest algorithm consistently delivers high accuracy in predicting hepatitis and related liver conditions achieving over 90% accuracy in many datasets [2], [13], [16]. Meanwhile, light gradient boosting machine (LightGBM) has emerged as a competitive alternative, outperforming other models on benchmark datasets such as Indian liver patient dataset (ILPD) [17]–[19]. While linear regression is frequently used as a baseline model in medical studies, it tends to perform less accurately than non-linear models such as random forest or boosting methods [4], [20].

This study aims to predict survival outcomes in hepatitis patients by comparing the performance of three widely used machine learning algorithms: linear regression, random forest, and LightGBM. The dataset includes public data from the UCI machine learning repository and real-world medical records collected from hospitals in Ambon city, Maluku–Indonesia. The goal is to identify the algorithm with the highest prediction accuracy and determine the most influential factors affecting patient survival, particularly within the Indonesian context [21]–[23].

2. MATERIALS AND METHOD

A wide range of research has been conducted to predict mortality rates and survival outcomes in hepatitis cases using machine learning and artificial intelligence (AI) approaches, aiming to optimize model performance for real-world applicability [10], [24]–[27]. These studies apply diverse machine learning methods across multiple hepatitis types (A, B, C, D, E), using structured datasets for both clinical and demographic features [5], [28]. Several algorithms have been deployed in hepatitis research, including logistic regression [12], random forest and naïve Bayes, as well as hybrid models such as improved random forests with support vector machines (SVMs) [29]. Some studies have extended into life expectancy prediction using K-nearest neighbors (KNN), enhanced with genetic algorithms, demonstrating the expansive exploration of algorithmic solutions in this domain [6], [30], [31].

This study evaluates three commonly used machine-learning techniques. These include a linear-based model (linear regression), a tree-ensemble method (random forest), and a gradient-boosting framework known as LightGBM. The novelty lies in evaluating their comparative performance in predicting hepatitis patient survival outcomes based on real-world and benchmark datasets. Understanding the theoretical foundations and strengths of these methods is essential for justifying their selection and interpreting results.

2.1. Linear regression

Linear regression serves as a fundamental statistical approach for exploring how predictor variables contribute to variations in an outcome variable. It is frequently used as a baseline algorithm in clinical data modeling due to its interpretability and simplicity [27], [32]. Despite its limitations in handling non-linear relationships, its inclusion in this study allows for comparison against more complex models.

2.2. Random forest

Random forest operates by aggregating the outputs of numerous decision trees, enabling the model to generalize effectively across heterogeneous clinical features. It reduces variance by averaging results across trees and is known for its robustness in handling noisy data, imbalanced classes, and high-dimensional datasets [2], [13], [16], [33]. Random forest has consistently demonstrated strong predictive performance in hepatitis and liver disease-related studies [17], [18], [34].

2.3. LightGBM

LightGBM applies gradient-boosted decision trees to learn complex patterns efficiently, offering faster training and strong performance on structured medical data. It is designed to be distributed and efficient, with faster training speed, lower memory usage, and better accuracy compared to traditional boosting methods [24], [26]. LightGBM has shown excellent results in biomedical datasets, including ILPD and hepatitis data, and is capable of handling large-scale, high-dimensional data efficiently [18], [19], [35].

2.4. Classification performance measurement

To evaluate the effectiveness of classification models, robust performance metrics are essential. In this study, a confusion matrix is used to measure accuracy, precision, recall, and F1-score by comparing the predicted classifications with the actual outcomes. This metric is effective in both binary and multi-class

classification problems and is widely used in medical prediction research [7], [12]–[14]. These performance metrics are based on four different combinations of predicted and actual values. Further explanation shown in the Table 1.

Table 1. Confusion matrix

Prediction result	Real situation	
	Positive class	Negative class
Positive class	TP	FP
Negative class	FN	TN

- True positive (TP) is the number of correct predictions on data whose actual value is also true.
- False negative (FN) occurs when data that should be classified as positive is mistakenly predicted as negative by the model. This means the model fails to identify positive data and incorrectly classifies it as negative.
- False positive (FP) It is a condition of the actual data that is wrong (negative data) but is predicted as true data.
- True negative (TN) That is, the prediction is correct as negative data according to the actual data condition is true as negative data.

To evaluate the overall performance of the model's predictions, accuracy metrics are employed. The accuracy score is calculated using a standard formula derived from the elements of the confusion matrix, as presented in Table 1, using the following (1).

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

Where TP denotes true positive, TN is true negative, FP is false positive, and FN is false negative. Additionally, this chapter outlines the research methodology applied in the study. In general, the research process consists of several key stages, which are illustrated in Figure 1.

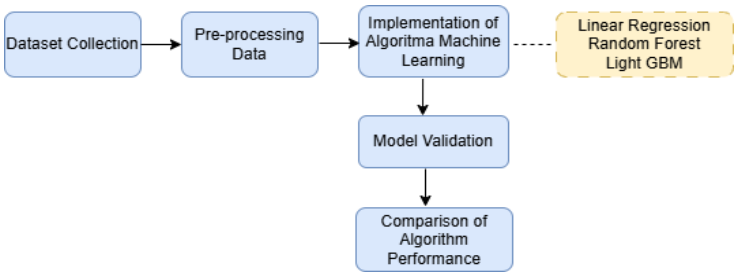

Figure 1. Research stages

- Dataset collection: in this stage, the datasets required for the study are gathered. These datasets include comprehensive information on medical history, laboratory test results, and diagnostic data related to liver health. The primary dataset used in this research is obtained from the UCI machine learning repository [36]. In addition, real-world clinical data were collected through direct field studies at several hospitals in Ambon city, Maluku, Indonesia.
- Data preprocessing: this phase involves cleaning the data to remove noise and inconsistencies, normalizing values, and eliminating redundant or irrelevant entries. Feature selection is also conducted to remove attributes that do not significantly contribute to the classification and prediction processes. This ensures that the dataset is consistent and suitable for the machine learning algorithms to be applied.
- Implementation of machine learning algorithms: once the dataset has been preprocessed, it is split into two subsets: training and validation/testing. This division allows for model optimization during training and performance evaluation during validation. This research employs three categories of predictive modelling techniques: a linear-based method represented by linear regression, a tree-ensemble strategy exemplified by random forest, and an advanced gradient-boosting framework commonly known as LightGBM. All implementation procedures are conducted using Google Colab as the computational environment.

- Model validation: this stage involves validating the trained models using the validation dataset. The performance of each algorithm is assessed based on accuracy metrics, which serve as indicators of prediction reliability. Each model is evaluated using the same validation protocol to ensure fair comparison.
- Algorithm performance comparison: in the final stage, the performance of all three algorithms is compared. After obtaining the accuracy metrics from the validation phase, a comparative analysis is performed to identify the algorithm with the most reliable and accurate predictive capabilities. This analysis supports the selection of the most effective model for hepatitis survival prediction.

3. RESULTS AND DISCUSSION

3.1. Dataset collection

This research utilized a dataset obtained from the UCI machine learning repository [12], [36], which was supplemented with original clinical data collected from various hospitals and health facilities in Ambon city, Maluku, Indonesia. A total of 154 patient records were used, each containing 19 independent variables, including clinical symptoms and laboratory test results relevant to hepatitis diagnosis. The independent variables included: Age, Sex, Steroids, Antiviral, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Alk Phosphate, SGOT, Albumin, Protime, and Histology. The dependent variable was the survival status of each hepatitis patient, labeled as either “Live” or “Die”. The selected features were chosen based on their clinical relevance to hepatitis progression and prognosis [5], [6].

3.2. Data preprocessing

The statge of preprocessing involved cleaning the dataset by handling missing values, correcting inconsistent data types, and removing duplicate entries. Categorical variables were transformed into numerical format to suit the machine learning algorithms. An exploratory data analysis (EDA) was also performed, including correlation analysis between features to assess inter-variable relationships.

In general, from the results of data exploration, it is known that there is a positive correlation in the variables ‘bilirubin’ and ‘alk_phosphate’ shown in Figure 2. The greater the value, the greater the positive correlation shown. This correlation is important to see the extent of the relationship between variables in the data. After all preprocessing procedures, the dataset was separated into two segments, where the larger segment supported model training and the smaller segment served for testing and evaluation.

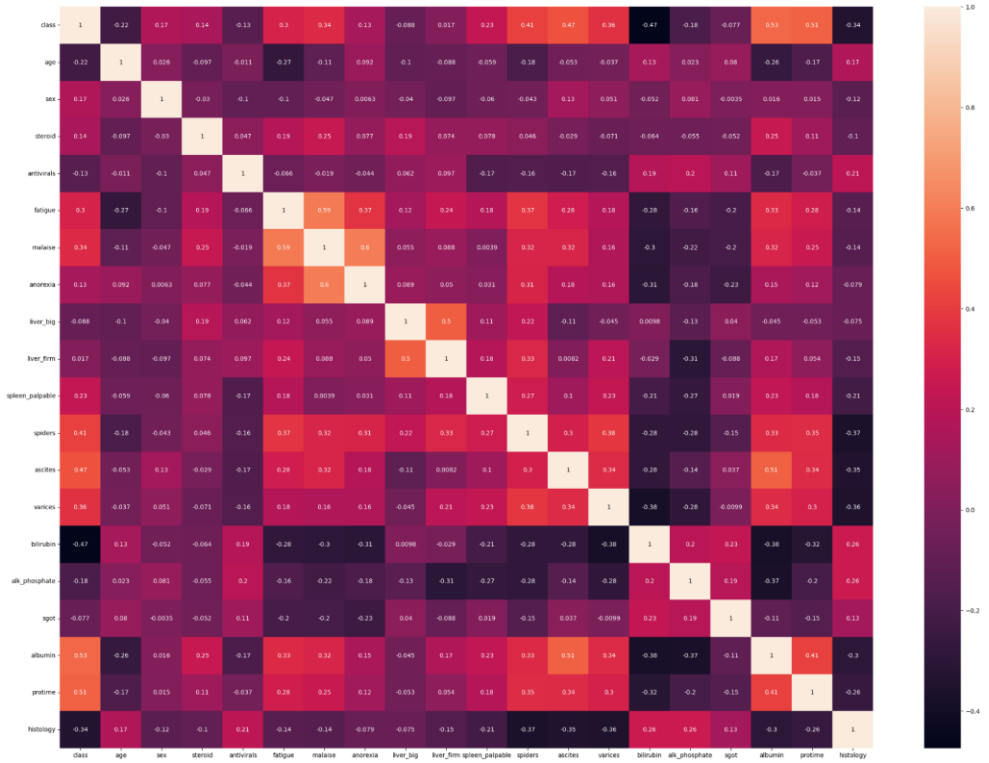

Figure 2. Correlation of various independent variables in the dataset using Heatmap

3.3. Implementation of machine learning algorithms

Three machine learning models were implemented: linear regression, random forest, and LightGBM. The cleaned dataset was trained and tested on each model to assess its ability to classify the survival status of hepatitis patients into two classes: live and die. The processes of training and evaluating the models were executed in an online computing environment, with Google Colab serving as the main platform. Each algorithm was trained using identical data splits and evaluation criteria to ensure fair comparison.

3.4. Model validation

To evaluate model performance, classification results were analyzed using confusion matrices and several evaluation metrics: accuracy, precision, recall, and F1-score. The confusion matrix for each algorithm is presented below on Table 2. The interpretation of Table 2 illustrates how the model categorized the test samples, detailing the distribution of correct and incorrect predictions across all four outcome types. As many as 20% of the total dataset, namely 154 data, 31 data were used as test sets and provided prediction results for positive and correctly predicted data conditions as many as 8 data, 14 data were correctly predicted but predicted incorrectly by the model, 7 data were incorrectly predicted (as positive data) and 2 data conditions were negatively predicted (actual data was wrong). According to the results of the prediction, the accuracy level obtained by the regression linear learning model is 0.322580 or 32%.

Table 2. Confusion matrix–linear regression

Prediction result	Real situation	
	Positive class	Negative class
Positive class	8	7
Negative class	14	2

In Table 3, the classification based on confusion matrix also shows the prediction results for 20% of the test data from the total data owned. The data condition is correct and predicted correctly by this learning model as many as 16 data, the data is correct but predicted incorrectly as many as 5, the data condition is wrong and predicted correctly 0 data and the data condition is incorrectly predicted as incorrect data as 10 data. From the results of this prediction, the accuracy level obtained by the random forest learning model is 0.838709 or 84%.

According to the classification results in Table 4, it can be seen that the model successfully predicted the data correctly for positive data as many as 14 data, the correct data and predicted wrong data by the model as many as 7 data, the wrong data and predicted as true data 0 data and the wrong data (negative) data predicted correctly as wrong data as many as 10 data. This shows the level of accuracy obtained by the learning model, which is 0.7741 or 77%.

Table 3. Confusion matrix–random forest

Prediction result	Real situation	
	Positive class	Negative class
Positive class	16	0
Negative class	5	10

Table 4. Confusion matrix–LightGBM

Prediction result	Real situation	
	Positive class	Negative class
Positive class	8	7
Negative class	14	2

3.5. Algorithm performance comparison

Based on the classification and prediction results obtained from the learning models, linear regression, the comparative assessment reveals that random forest outperformed the other models, attaining an accuracy of 84%. LightGBM achieved 77%, and linear regression showed the weakest performance with 32% accuracy. These findings support earlier studies [1], [2], which also highlighted the strong predictive performance of random forest in liver disease classification. However, this study goes further by combining reference data with real-world clinical data collected from actual healthcare settings. This integration provides a more localized and realistic view of how the models perform in practice, especially in environments where variability and data quality often differ from controlled research datasets.

Looking at the accuracy results through the confusion matrix, random forest consistently delivered the most accurate predictions, placing it at the top, followed by LightGBM. linear regression, by contrast, lagged behind, and this may be due to how the model operates differently from the other two algorithms. In the case of random forest and LightGBM, classification and prediction processes were applied directly to the models. But with linear regression, a conversion step was needed to turn continuous outputs into binary form before the model could be evaluated for classification tasks. This not only adds an extra layer of complexity but also exposes one of the model's main weaknesses, its limited ability to handle binary clinical classification, especially when working with non-linear data like hepatitis progression.

This underscores the importance of selecting algorithms that are not only accurate but also well-matched to the structure and characteristics of the data. In this research, the random forest model is clearly the most effective among those evaluated. However, the performance of the model is not solely dependent on the algorithm selection but is also significantly influenced by factors such as the dataset size, the relevance of the features, and the quality of the input data. Going forward, further evaluation of other models using larger and more diverse datasets would be valuable to better understand the robustness and generalizability of each learning approach. In addition, the findings point to the promising role of ensemble-based algorithms, particularly random forest, as practical tools in intelligent clinical decision-support systems for early detection and treatment planning of hepatitis.

3.6. Clinical insights

Clinical experts emphasized that hepatitis viruses are classified into five major forms, which differ in how they spread, how they present clinically, and the severity of their mortality risk. Accurate classification is essential in guiding diagnostic and treatment decisions. Moreover, understanding transmission pathways and patient behaviors is crucial for prevention, reinforcing the importance of hygiene and dietary management in mitigating hepatitis transmission risks.

4. CONCLUSION

This work applied and assessed three predictive approaches, linear regression, random forest, and LightGBM, to estimate survival outcomes in hepatitis cases. The comparative results indicate that random forest delivered the strongest performance with 84% accuracy, LightGBM attained 77%, and linear regression showed the weakest result at 32%. These results are significant because they validate the applicability of ensemble learning models, particularly random forest, in clinical prediction tasks using real-world patient data. Compared to existing research, this study contributes a context-specific model tailored to healthcare conditions in Ambon, Indonesia, bridging the gap between theoretical models and field applicability. The lower performance of linear regression reinforces the importance of algorithm selection based on data characteristics and the nature of the prediction task.

Ultimately, these findings demonstrate that the random forest algorithm offers an accurate and adaptive solution for predicting survival in hepatitis cases, especially when trained using real-world medical data. Its performance demonstrates that this algorithm has strong potential for integration into intelligent healthcare systems, particularly in resource-limited settings. For future research, several improvements are suggested, including expanding the dataset size to reduce the risk of overfitting and improve generalizability, evaluating additional machine learning algorithms such as deep learning approaches to explore further performance gains, and classifying predictions based on hepatitis types (A, B, C, and D) to enable more granular and disease-specific prognostic models. These enhancements are expected to contribute to the development of more accurate, reliable, and clinically applicable decision-support systems for hepatitis diagnosis and prognosis.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Research Institute of Universitas Kristen Indonesia Maluku for the support and facilities provided throughout the course of this study. Special thanks are also extended to RSUP Dr. J. Leimena, RS Sumber Hidup and other hospitals in Ambon city, Maluku, for their collaboration and assistance in providing access to real-world clinical data.

FUNDING INFORMATION

This research was funded by the Research Institute of Universitas Kristen Indonesia Maluku through the UKIM Flagship Research Program, fiscal year 2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hennie Tuhuteru	✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	✓
Goldy Valendria Nivaan	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Marvelous Marvel Rijoly	✓	✓	✓	✓		✓		✓		✓	✓	✓		
Joselina Tuhuteru	✓	✓		✓	✓	✓	✓			✓		✓		

C : Conceptualization
M : Methodology
So : Software
Va : Validation
Fo : Formal analysis
I : Investigation
R : Resources
D : Data Curation
O : Writing - Original Draft
E : Writing - Review & Editing
Vi : Visualization
Su : Supervision
P : Project administration
Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, HT, upon reasonable request.

REFERENCES

[1] S. M. Ganie and P. K. D. Pramanik, "Predicting chronic liver disease using boosting technique," *International Conference on Artificial Intelligence for Innovations in Healthcare Industries, ICAIIHI 2023*, 2023, doi: 10.1109/ICAIIHI57871.2023.10489026.

[2] K. M. Almustafa and J. Katrib, "Evaluating machine learning classifiers for predicting liver disease outcomes," *2024 International Conference on Decision Aid Sciences and Applications, DASA 2024*, 2024, doi: 10.1109/DASA63652.2024.10836587.

[3] M. M. Majzoobi, S. Namdar, R. Najafi-Vosough, A. A. Hajilooi, and H. Mahjub, "Prediction of hepatitis disease using ensemble learning methods," *Journal of preventive medicine and hygiene*, vol. 63, no. 3, pp. E424–E428, 2022, doi: 10.15167/2421-4248/jpmh2022.63.3.2515.

[4] N. K. Kumar and D. Vigneswari, "Hepatitis- infectious disease prediction using classification algorithms," *Research Journal of Pharmacy and Technology*, vol. 12, no. 8, pp. 3720–3725, 2019, doi.org/10.5958/0974-360X.2019.00636.X

[5] Y. Guo, Y. Feng, F. Qu, L. Zhang, B. Yan, and J. Lv, "Prediction of hepatitis e using machine learning models," *PLoS ONE*, vol. 15, no. 9 September, 2020, doi: 10.1371/journal.pone.0237750.

[6] J. Yang, "Hepatitis C risk prediction based on adaboost," *Highlights in Science, Engineering and Technology*, vol. 54, pp. 413–419, 2023, doi: 10.54097/hset.v54i.9803.

[7] R. K. Sachdeva, P. Bathla, P. Rani, V. Solanki, and R. Ahuja, "A systematic method for diagnosis of hepatitis disease using machine learning," *Innovations in Systems and Software Engineering*, vol. 19, no. 1, pp. 71–80, 2023, doi: 10.1007/s11334-022-00509-8.

[8] M. A. Hezari, M. Baes, A. A. Hezari, and M. Hassanbabaei, "Advanced predictive modeling for hepatitis C diagnosis using machine learning," *Clinical and Molecular Epidemiology*, vol. 1, p. 12, 2024, doi: 10.53964/cme.2024012.

[9] H. D. Saputra, A. I. E. Efendi, E. Rudini, D. Riana, and A. S. Hewiz, "Hepatitis prediction using K-NN, naive bayes, support vector machine, multilayer perceptron and random forest, gradient boosting, K-Means," *Journal Medical Informatics Technology*, pp. 96–100, 2023, doi: 10.37034/medinftech.v1i4.21.

[10] A. Alizargar, Y. L. Chang, and T. H. Tan, "Performance comparison of machine learning approaches on hepatitis c prediction employing data mining techniques," *Bioengineering*, vol. 10, no. 4, 2023, doi: 10.3390/bioengineering10040481.

[11] World Health Organization (WHO), *Global hepatitis report 2024 Action for access in low- and middle-income countries*. 2024.

[12] G. V. Nivaan and A. W. R. Emanuel, "Analytic predictive of hepatitis using the regression logic algorithm," *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, pp. 106–110, 2020, doi: 10.1109/ISRITI51436.2020.9315365.

[13] R. Y. Krisnabayu, A. Ridok, and A. S. Budi, "Hepatitis detection using random forest based on SVM-RFE (recursive feature elimination) feature selection and SMOTE," *ACM International Conference Proceeding Series*, pp. 151–156, 2021, doi: 10.1145/3479645.3479668.

[14] A. Dhyani *et al.*, "Comparative analysis of supervised machine learning algorithms for liver disease prediction with SMOTE Enhancement," *2023 3rd Asian Conference on Innovation in Technology, ASIANCON 2023*, 2023, doi: 10.1109/ASIANCON58793.2023.10270381.

[15] M. J. Nayeem, S. Rana, F. Alam, and M. A. Rahman, "Prediction of hepatitis disease using K-Nearest neighbors, naive bayes, support vector machine, multi-layer perceptron and random forest," *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, pp. 280–284, 2021, doi: 10.1109/ICICT4SD50815.2021.9397013.

[16] M. Arif, M. Abbas, M. A. Shehzad, Z. Batool, M. Rabia, and A. M. Soomro, "An ensembling approach to predict hepatitis in patients with liver disease using machine learning," *VFAST Transactions on Software Engineering*, vol. 11, no. 3, pp. 42–52, 2023, doi: 10.21015/vtse.v11i3.1598.
[17] S. M. Ganie and P. K. Dutta Pramanik, "A comparative analysis of boosting algorithms for chronic liver disease prediction," *Healthcare Analytics*, vol. 5, 2024, doi: 10.1016/j.health.2024.100313.
[18] U. N. Yefou, P. O. M. Choudja, B. Sow, and A. Adejumo, "Optimized machine learning models for hepatitis c prediction: leveraging optuna for hyperparameter tuning and streamlit for model deployment," *Communications in Computer and Information Science*, vol. 2068 CCIS, pp. 88–100, 2024, doi: 10.1007/978-3-031-57624-9_5.
[19] Y. Wang, B. Yin, and Q. Zhu, "Application of machine learning algorithms in predicting hepatitis C," *ACM International Conference Proceeding Series*, pp. 359–365, 2023, doi: 10.1145/3644116.3644176.
[20] S. Lakumarapu, R. Nithyanandhan, V. S. Bhargavi, T. P. Anish, M. Nalini, and R. Siva Subramanian, "Machine learning approaches for liver disease prediction: a comparative analysis," *5th International Conference on Electronics and Sustainable Communication Systems, ICESC 2024 - Proceedings*, pp. 159–164, 2024, doi: 10.1109/ICESC60852.2024.10689974.
[21] S. S. Nigatu, P. C. R. Alla, R. N. Ravikumar, K. Mishra, G. Komala, and G. R. Chami, "A comparitive study on liver disease prediction using supervised learning algorithms with hyperparameter tuning," *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023*, pp. 353–357, 2023, doi: 10.1109/InCACCT57535.2023.10141830.
[22] C. Raiaramesh, R. Nayak, O. S. Naaesh, and P. L. Kanth, "Liver disease prediction using machine learning algorithms with comparative analysis of different algorithms," *2023 2nd International Conference on Ambient Intelligence in Health Care, ICAIHC 2023*, 2023, doi: 10.1109/ICAIHC59020.2023.10431470.
[23] D. F. Santos, "Predicting the severity of hepatitis c using machine learning models," 2023, [Online]. Available: https://orcid.org/0000-0002-8599-9436
[24] E. Ramadanti, D. A. Dinathi, C. Christianskaditya, and D. R. Chandranegara, "Diabetes disease detection classification using light gradient boosting (LightGBM) with hyperparameter tuning," *Sinkron*, vol. 8, no. 2, pp. 956–963, 2024, doi: 10.33395/sinkron.v8i2.13530.
[25] H. M. Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C virus prediction based on machine learning framework: a real-world case study in Egypt," *Knowledge and Information Systems*, vol. 65, no. 6, pp. 2595–2617, 2023, doi: 10.1007/s10115-023-01851-4.
[26] D. Zhang and Y. Gong, "The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure," *IEEE Access*, vol. 8, pp. 220990–221003, 2020, doi: 10.1109/ACCESS.2020.3042848.
[27] D. Haryadi, D. M. U. Atmaja, and A. R. Hakim, "Prediction of liver disease using a linear regression algorithm," *Journal of Informatics and Communication Technology (JICT)*, vol. 5, no. 1, pp. 89–100, 2023, doi: 10.52661/j_ict.v5i1.182.
[28] T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu, and T. Islam, "Detection of hepatitis (A, B, C and E) viruses based on random forest, k-nearest and naïve bayes classifier," *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, 2019, doi: 10.1109/ICCCNT45670.2019.8944455.
[29] U. K. Lilhore *et al.*, "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-36605-3.
[30] A. M. Ali *et al.*, "Explainable machine learning approach for hepatitis C diagnosis using SFS feature selection," *Machines*, vol. 11, no. 3, 2023, doi: 10.3390/machines11030391.
[31] A. M. Al Alawi, H. H. Al Shuaili, K. Al-Naamani, Z. Al Naamani, and S. A. Al-Busafi, "A machine learning-based mortality prediction model for patients with chronic hepatitis C infection: an exploratory study," *Journal of Clinical Medicine*, vol. 13, no. 10, 2024, doi: 10.3390/jcm13102939.
[32] P. Schober and T. R. Vetter, "Linear regression in medical research," *Anesthesia and Analgesia*, vol. 132, no. 1, pp. 108–109, 2021, doi: 10.1213/ANE.0000000000005206.
[33] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: a review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.
[34] T. R. P. Lestari, "Global collaboration in hepatitis management: Indonesia's position and role (In indonesian)," *Bidang Kesejahteraan Rakyat Info Singkat*, pp. 21–20, 2024.
[35] M. W. Ali, A. Gupta, M. Khan, and M. Wajid, "Non-contact breath rate classification using SVM model and mmWave radar sensor data," in *Proceedings of the 2024 2nd International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles, ICPEEV 2024*, 2024. doi: 10.1109/ICPEEV63032.2024.10931988.
[36] A. Asuncion and D. J. Newman, "UCI machine learning repository: data sets," University of California Irvine School of Information, 2007.

BIOGRAPHIES OF AUTHORS

Hennie Tuhuteru is Vice Dean I, academic field of the Faculty of Computer Science, Universitas Kristen Indonesia Maluku (UKIM). His research areas are information system, web development, big data, and data science. He previously pursued a master's degree at Universitas Kristen Satya Wacana (UKSW) Salatiga, became a lecturer and wrote many publications related to his field of knowledge. He previously served as Head of the Quality Assurance Unit of the Faculty of Computer Science for approximately 5 years. Currently, he serves as Vice Dean I of the Faculty of Computer Science, UKIM. He also has participated in many scientific research and services. In recent years, he has even qualified for national funding for research and community service. He can be contacted at email: hannytuhuteru@gmail.com or hennietuhuteru@ukim.ac.id.

Goldy Valendria Nivaan received an S.Kom. degree in information systems at Nusamandiri University Jakarta and an M.Kom. degree in informatics at Atma Jaya University Yogyakarta. Despite her youth, she is currently a lecturer at the Department of Informatics and Vice Dean II of the Faculty of Computer Science, Universitas Kristen Indonesia Maluku (UKIM). She has mentored and co-supervised more than 20 students and has authored or co-authored several publications as well as participated in various scientific writing training activities both nationally and internationally that support the writing of scientific papers. Her research interests include soft computing, machine learning, and intelligent systems. She can be contacted at email: valendria17@gmail.com.

Marvelous Marvel Rijoly his higher studies in information system department, Universitas Kristen Satya Wacana (UKSW). He works at the Computer Laboratory of the Faculty of Computer Science, Universitas Kristen Indonesia Maluku (UKIM). To support his interest in the academic field, in this case scientific research, he has attended the data science bootcamp for approximately 9 months. This activity provides training that supports data pre-processing, EDA, machine learning, and computer vision. He can be contacted at email: marvelmarvin1692@gmail.com.

Joselina Tuhuteru completed her master's degree in management, Universitas Kristen Satya Wacana (UKSW), Salatiga in 2014. She focuses on management and financial behavior. She is currently the Head of External Quality Assurance System (SPME), quality assurance institute of Universitas Kristen Indonesia Maluku (UKIM). In the last four years she has produced much research in accordance with his field both nationally funded by the directorate of research, technology, and community service, internally at the University and independently. She can be contacted at email: joselina.tuhuteru@gmail.com.