Jurnal TRANSFORMATIKA
Vol.22, No.1, Juli 2024, pp. 61 - 71
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Bootstrapped Aggregating Optimization in Random Forest for Hepatitis Risk

Marselina Endah Hiswati[1*], Mohammad Diqi[2], Endang Nurul Syafitri[3], Annur Fauziyyah[4]
[1]Department of Informatics, Universitas Respati Yogyakarta
Jl. Laksda Adisucipto Km 6.3 Depok Sleman, e-mail: marsel.endah@respati.ac.id
[2] Department of Informatics, Universitas Respati Yogyakarta
Jl. Laksda Adisucipto Km 6.3 Depok Sleman, e-mail: diqi@respati.ac.id
[3] Department of Nursing, Universitas Respati Yogyakarta
Jl. Raya Tajem Km 1.5 Depok Sleman, e-mail: e.nurul.s@respati.ac.id
[4] Department of Informatics, Universitas Respati Yogyakarta
Jl. Laksda Adisucipto Km 6.3 Depok Sleman, e-mail: 22220021@respati.ac.id

ARTICLE INFO

History of the article :
Received 24 April 2024
Received in revised form 30 Juni 2024
Accepted 21 Juli 2024
Available online 31 Juli 2024

Keywords:

Hepatitis; Random Forest; Bootstrapped Aggregating; Risk Prediction; Ensemble Techniques

*** Correspondence:**

Telepon:
+62 838 4030 2028

E-mail:
marsel.endah@respati.ac.id

ABSTRACT

This research optimizes the Random Forest model with Bootstrapped Aggregating to predict hepatitis risk. The global significance of hepatitis as a health problem is underscored by its widespread impact. Using a Kaggle dataset comprising 596 records and 20 attributes, including age categories and gender, the study identifies limitations in predicting hepatitis risk. Through hyperparameter optimization, such as adjusting the number and depth of trees, the Random Forest model with bootstrapped aggregate achieves an accuracy of 96%, surpassing the standard model's 88%. The results demonstrate a significant improvement in precision, recall, and f1 score, particularly in reducing false negatives. The conclusion highlights the practical potential of this model for a more accurate assessment of hepatitis risk. While acknowledging limitations related to the size of the dataset, these findings provide a foundation for developing predictive models in the context of hepatitis risk, emphasizing the importance of employing ensemble techniques to improve model performance.

1. INTRODUCTION

Hepatitis, particularly in types B and C, is a formidable challenge for health systems worldwide [1]. The World Health Organization describes hepatitis as a liver disease characterized by inflammation that arises from various etiologies, including viral infections, exposure to harmful chemicals, or the body's immune system turning against its liver tissue [2]. This disease presents a critical threat to the collective health of populations due to its infectious nature and propensity to trigger persistent health issues, including chronic liver disease and other long-term complications, which underscore its potential to dramatically impact public health on a global scale [3].

The worldwide incidence of hepatitis has emerged as a pressing issue, given its association with a spectrum of severe health conditions such as liver cirrhosis, end-stage liver failure, and a higher likelihood of liver carcinoma [4]. These infections carry weighty economic repercussions, evidenced by substantial healthcare expenditures, loss of workforce participation due to illness-related absences, and downturns in overall productivity levels [5]. Furthermore, the persistent propagation of hepatitis is a formidable barrier to the fruition of critical international health benchmarks set by the Sustainable Development Goals—most notably, SDG 3, which is dedicated to fostering optimal health and well-being for the entire global populace [6].

The identified research problem in this study is associated with limitations or deficiencies in the quality of hepatitis risk prediction [7]. Although predicting hepatitis risk is a critical element in preventing and managing this disease, specific aspects still require further attention to enhance the accuracy and reliability of such predictions [8]. The identification and deeper understanding of these limitations are expected to pave the way for developing more effective prediction methods, thereby strengthening overall efforts in preventing and intervening in hepatitis.

This study aims to optimize the Bootstrapped Aggregating parameters in the Random Forest algorithm for predicting hepatitis risk. With a focus on advancing prediction techniques, this research seeks to enhance the accuracy and reliability of the model in predicting hepatitis risk. Adjusting the Bootstrapped Aggregating parameters in the Random Forest algorithm is expected to identify an optimal configuration that can produce more accurate prediction results, assist in early identification of hepatitis risk, and improve the effectiveness of prevention and intervention strategies in managing this disease.

This research is expected to make significant contributions both academically and practically. Academically, the study aims to enrich the literature by improving our understanding of the factors influencing the prediction of hepatitis risk. By optimizing the Bootstrapped Aggregating parameters in the Random Forest algorithm, this research can provide new insights into more effective prediction methods within the healthcare context. Practically, the research findings are anticipated to serve as a foundation for developing more reliable models for predicting hepatitis risk, offering tangible benefits in supporting clinical decisions, and designing more targeted and efficient prevention and intervention strategies in the clinical management of hepatitis.

LITERATURE REVIEW

Risk prediction methods for hepatitis in prior studies have tackled the intricate interactions among risk factors and accounted for the variability in clinical conditions. Various machine learning algorithms such as Support Vector Machine, K-Nearest Neighbor, and Artificial Neural Networks have been applied to classify and predict hepatitis [9]. Naïve Bayes Classifier, Logistic Regression, and J48 Decision Tree have been employed for classification, alongside filter-based feature selection techniques such as Cfs Subset Eval, Data Gain Attribute Eval, and Principal Components [10]. A deep learning-driven decision support system utilizing bidirectional long/short-term memory (BiLSTM) has also been suggested for precise binary classification of hepatitis diagnosis [11]. Fuzzy logic has been used to manage imprecise and intricate risk factors,

aligning them with hepatitis as the output variable [12]. Random Forest has exhibited high accuracy in hepatitis diagnosis and feature selection [13]. These methodologies acknowledge the complexity of interactions among risk factors and the variability in clinical conditions, thereby enhancing hepatitis risk prediction.

Previous hepatitis risk prediction techniques, particularly machine learning algorithms, have shown advantages and disadvantages. The benefits include the ability to analyze many clinical parameters with speed and accuracy, leading to improved identification of factors influencing patient survival rates [14]. Machine learning techniques such as SVM and XGBoost have demonstrated high accuracy and AUC in predicting hepatitis C, making them practical tools for early diagnosis and treatment [15]. However, there are also limitations to these approaches. Some studies have reported lower accuracy rates, such as 72%, when using decision tree models for hepatitis C prediction [16].

Moreover, the effectiveness of various machine learning methods can differ, with SVM surpassing KNN in accuracy, error rate, specificity, and negative prediction value [17]. The findings of this research enhance understanding by pinpointing particular clinical parameters like LIVER BIG, LIVER FIRM, SPLEEN PALPABLE, and ANOREXIA that enhance patient survival rates [18]. Additionally, the study underscores the potential of the Voting classifier model in accurately predicting liver disease occurrence, as evidenced by its high accuracy and AUC.

Previous research has endeavored to enhance the efficacy of Random Forest forecasts through parameter adjustments such as the number of trees (n_estimators) and tree depth (max_depth). One approach, the Reducing and Aggregating Random Forest Trees by Elastic Net (RARTEN) method, suggests employing the random forest algorithm for prediction, utilizing Elastic Net regression to decrease the tree count, and then amalgamating the chosen trees [19]. Another study observed that decisions regarding parameters such as training/testing data partitioning strategies, variable selection, and the ratio of training to testing data significantly impacted the goodness-of-fit of Random Forest models [20]. In software defect prediction, a Multi-Objective Random Forest (MO-RF) algorithm was introduced with a data resampling technique to tackle class imbalance problems, demonstrating enhanced performance compared to other prediction models [21]. Furthermore, a study aimed to boost classification accuracy by converting continuous attributes into categories and discovered that Random Forest classification with continuous attribute transformation outperformed the original dataset model in certain training data variations [22]. Another research effort proposed an enhanced algorithm that merges feature fusion and random forests quantile classifier to improve the overall classification performance of the classifier for unbalanced data [23].

Previous studies have delved into specific techniques for processing data and features to enhance the quality of Random Forest predictions. One research effort concentrated on converting continuous attributes into categories by generating percentile values randomly as thresholds for categorization. The study assessed four algorithms for generating these percentile values and chose the optimal model based on minimal variability and the distribution of revenue expectations [22]. Another study explored feature selection to enhance the performance of Random Forest models for land use and land cover mapping. This study employed a Random Forest-based feature selection method utilizing Sentinel-1, -2, and Shuttle Radar Topographic Mission (SRTM) data, concluding that integrating these data sources improved classification accuracy compared to using Sentinel-2 data alone [24].

Furthermore, a study investigated the impact of parameter choices in Random Forest models for forecasting nitrate concentrations in aquatic environments. It was discovered that considering temporal dependencies during data splitting and optimizing variable selection was crucial for model fitting and accuracy [20]. Another study combined Information Gain, Fast Fourier Transform (FFT), and Synthetic Minority Oversampling Technique (SMOTE) methods to enhance

DOI : https://10.26623/transformatika.v22i1.9073

feature selection in Random Forest models, leading to improved performance accuracy [25]. Lastly, an enhanced algorithm that combines feature fusion and a random forests quantile classifier was suggested to tackle issues related to low prediction performance and imbalanced data. This algorithm improved the overall classification performance of samples for imbalanced data compared to alternative algorithms [23].

Bootstrapped Aggregating (Bagging) parameter optimization in the Random Forest algorithm is applied to improve hepatitis risk prediction by comparing the performance of different methods. In a study by Majzoobi et al., traditional and ensemble learning methods, including bagging, random forest, AdaBoost, and logistic regression, were used to predict hepatitis B virus (HBV) and hepatitis C virus (HCV) [26]. The random forest method showed the best performance for predicting HBV, with ALT identified as the most essential variable [27]. For predicting HCV, random forest also had the highest accuracy, with AST, ALT, and age identified as essential variables [28]. The optimization of Bagging parameters in the Random Forest algorithm is expected to improve the accuracy and performance of the model in predicting hepatitis risk [29].

Bootstrapped Aggregating parameter adjustment, or bagging, can help address the potential for overfitting or underfitting in Random Forest models used for hepatitis risk prediction [26]. Bagging involves creating multiple subsets of the original dataset through resampling and training individual models on each subset. By aggregating the predictions of these models, the overall model becomes more stable and less prone to overfitting or underfitting. This technique improves the model's generalization ability and reduces the variance in predictions. In addition to bagging, other strategies can be implemented to enhance model stability and generalization in the datasets used for hepatitis risk prediction. These strategies include feature selection or screening to identify the essential variables [29], class balancing to address imbalanced datasets [30], and hyperparameter tuning to optimize the performance of the Random Forest model [31].

RESEARCH METHODS

1. Dataset and Preprocessing

The dataset used in this research was obtained from Kaggle. The dataset has 20 attributes, including the target attribute 'class' with values DIE and LIVE. Other attributes involve information related to the use of steroids, antivirals, fatigue, malaise, anorexia, and liver conditions, as well as several laboratory parameters and other patient characteristics. The dataset includes 596 records that will be utilized in the analysis of hepatitis risk prediction.

In the data preprocessing phase, a series of steps were taken to ensure the cleanliness and quality of the dataset. Data cleaning involves handling missing values and ensuring each attribute has valid values. Furthermore, data preparation involves converting categorical variables such as 'sex' into a format that the model can process. Finally, normalization is performed using MinMaxScaler to ensure that all attributes have a consistent range of values, avoiding the domination of specific attributes in the model. This preprocessing process aims to provide that the dataset is ready for use in developing the hepatitis risk prediction model.

2. Data Splitting

The dataset is divided into two main parts: the training and testing sets. Eighty percent of the total data is used as the training set to train the hepatitis risk prediction model. Meanwhile, the remaining 20% (120 samples) is isolated as the testing set, which will be used to test and evaluate the model's performance. This division aims to ensure that the model receives adequate training and is then tested on data it has never seen before to assess its general predictive capabilities.

3. Model Selection

Random Forest is a machine learning algorithm that leverages ensemble learning by combining multiple decision trees to enhance accuracy and reduce overfitting. Each tree is generated randomly by considering a small subset of the data and attributes. The prediction results from each tree are then aggregated to produce the final prediction. The main advantages of Random Forest include stability and its ability to handle large datasets with diverse attributes, making it a popular choice for various classification and regression problems.

Bootstrapped Aggregating, more commonly known as Bagging, is an ensemble learning technique involving the creation of several new datasets by resampling from the original dataset. Each of these new datasets is used to train identical models independently. Subsequently, the prediction results from each model are aggregated, often by taking the average, to generate the final prediction. Bagging effectively reduces variance and improves model stability, producing more consistent and accurate predictions.

Employing Bootstrapped Aggregating (Bagging) in the Random Forest algorithm is justified for multiple reasons. Firstly, Bagging helps mitigate overfitting and reduce variance by randomizing the creation of each decision tree, fostering a more stable and general model. Additionally, Bagging enhances the model's accuracy and resilience against dataset noise through resampling techniques. In predicting hepatitis risk, where stability and robustness are paramount, Bootstrapped Aggregating in Random Forest is expected to yield more consistent and reliable predictions. The model's alignment with the research objectives is evident in the hyperparameter definitions slated for optimization, encompassing the number of trees, tree depth, and minimum sample sizes for split and leaf. The optimization process seeks to find the optimal configuration that maximizes the model's ability to predict hepatitis risk, ensuring adaptability to the unique dataset characteristics employed in the study.

4. Evaluation Metrics

The confusion matrix assesses the model's performance by displaying the count of correct and incorrect predictions in a matrix format, comprising four cells: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). This matrix enables the calculation of evaluation metrics such as precision, recall, and F1-score.

Meanwhile, the classification report summarizes the model's performance by presenting precision, recall, and F1-score for each class in a classification task. Precision measures the accuracy of positive predictions, recall evaluates the model's ability to identify all positive instances, and the F1 score provides a balance between precision and recall. The corresponding formulas are shown in Equations 1-3.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

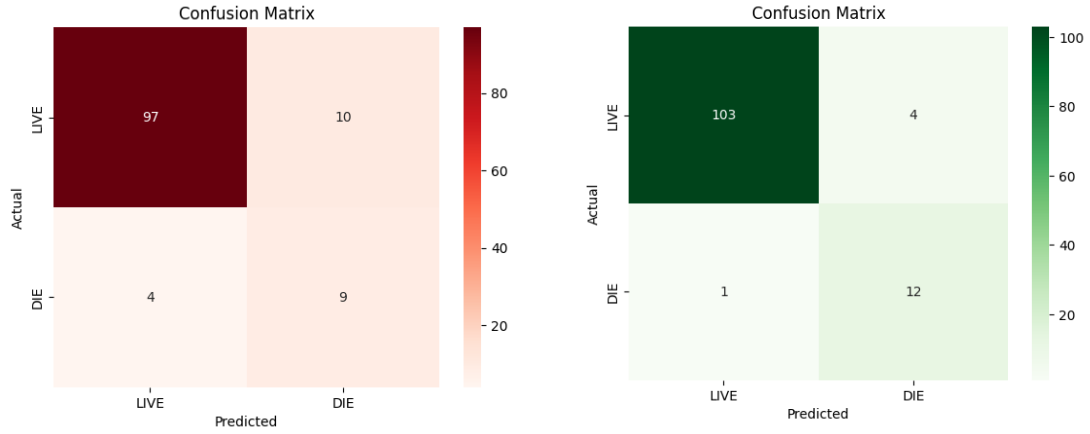
$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

By utilizing the confusion matrix and classification report, we can gain insights into how accurately and efficiently the model can make predictions, especially in hepatitis risk prediction research.

RESULTS

The research results reveal compelling insights into the predictive performance of the Random Forest algorithm in the context of hepatitis risk assessment. Two model variations were explored: one using standard Random Forest and the other incorporating Bootstrapped Aggregating (Bagging) for enhanced accuracy, as shown in Figure 1. The model's accuracy is a foundational metric, with the standard Random Forest achieving an accuracy of 88.00% and the version with Bootstrapped Aggregating demonstrating a notable improvement at 96.00%. These accuracy figures lay the groundwork for comprehensively examining the models' effectiveness in predicting hepatitis risk.



(a) (b)

Figure 1. Model's Performance (a) Standard Random Forest; (b) Random Forest with Bagging

The comparison of confusion matrices provides a detailed understanding of the model's performance. In the standard Random Forest, the confusion matrix exhibits 97 true positives, 10 false positives, 4 false negatives, and 9 true negatives. This configuration results in a commendable accuracy but reveals a notable proportion of false negatives, indicating instances where the model failed to identify actual positive cases. On the other hand, the Random Forest with Bootstrapped Aggregating demonstrates a superior confusion matrix, with 103 true positives, only 4 false positives, 1 false negative, and 12 true negatives. This configuration's reduced number of false positives and negatives signifies a more precise and reliable predictive capability, substantially improving overall accuracy. The utilization of Bootstrapped Aggregating in Random Forest has contributed to a model with enhanced discriminatory power, making it particularly promising for practical applications in hepatitis risk assessment.

Tables 1 and 2 below explore the classification report results of both models, discussing precision, recall, and f1-score for each class and commenting on the overall accuracy rates.

Table 1. Classification Report of Standard Random Forest

	precision	**recall**	**f1-score**	**support**
LIVE	0.96	0.91	0.93	107
DIE	0.47	0.69	0.56	13
accuracy			0.88	120
macro avg	0.72	0.80	0.75	120
weighted avg	0.91	0.88	0.89	120

Table 2. Classification Report of Random Forest with Bagging

	precision	**recall**	**f1-score**	**support**
LIVE	0.99	0.96	0.98	107
DIE	0.75	0.92	0.83	13
accuracy			0.96	120
macro avg	0.87	0.94	0.90	120
weighted avg	0.96	0.96	0.96	120

From the classification report of Random Forest with Bagging in Table 2, it can be observed that this model achieves an accuracy rate of 96%, with high precision for the 'LIVE' class (99%) and high recall for the 'DIE' class (92%). This indicates that the model can identify positive and negative cases. In contrast, the classification report for the standard Random Forest in Table 1 shows an accuracy of 88%, with lower precision for the 'DIE' class (47%) and suboptimal recall (69%). These results indicate that Bootstrapped Aggregating in Random Forest significantly improves precision and recall, resulting in a model with enhanced predictive capabilities. The Random Forest model with Bagging, exhibiting high precision and recall for both classes, demonstrates strong potential for more accurate hepatitis risk assessments.

DISCUSSION

1. Summarization of Key Findings

In addressing the research problem focused on optimizing the prediction of hepatitis risk, the study explored the efficacy of Random Forest models, both standard and enhanced with Bootstrapped Aggregating (Bagging). The primary goal was to improve the accuracy and reliability of hepatitis risk assessments. The key findings indicate a substantial enhancement in predictive performance by incorporating Bagging. The Random Forest with Bagging demonstrated a remarkable accuracy of 96%, surpassing the standard Random Forest, which achieved 88%. Noteworthy improvements were observed in precision and recall for both 'LIVE' and 'DIE' classes, reflecting the effectiveness of Bagging in reducing errors and enhancing the model's ability to accurately identify positive and negative cases. These findings underscore the significance of employing ensemble techniques, such as Bootstrapped Aggregating, in refining machine learning models for more precise hepatitis risk predictions.

DOI : https://10.26623/transformatika.v22i1.9073

2. Result Interpretations

The analysis of the results reveals discernible patterns and relationships within the data, shedding light on the effectiveness of the Random Forest models in predicting hepatitis risk. The observed improvement in accuracy, precision, and recall with the incorporation of Bootstrapped Aggregating aligns with expectations, demonstrating the efficacy of ensemble techniques in refining model predictions. The unexpected yet encouraging aspect significantly enhances the model's predictive capabilities, particularly in reducing false negatives. This unanticipated improvement suggests that including Bootstrapped Aggregating has a pronounced positive impact on the model's sensitivity, which is crucial for identifying positive cases. Possible alternative explanations for these results could involve the unique characteristics of the hepatitis dataset, emphasizing the importance of ensemble methods in mitigating biases and increasing the robustness of the model. Overall, the findings underscore the success of the chosen approach and highlight the potential for more accurate hepatitis risk assessments through thoughtful ensemble model design.

3. Research Implications

This research's implications extend to predictive modeling for hepatitis risk assessment, emphasizing the practical significance of incorporating ensemble techniques, specifically Bootstrapped Aggregating, into Random Forest algorithms. By substantially improving accuracy, precision, and recall, the study underscores the potential for enhancing the reliability of hepatitis risk predictions. These findings align with the existing literature, highlighting the advantages of ensemble methods in mitigating overfitting and increasing model robustness. The research contributes new insights by demonstrating the tangible benefits of Bagging in reducing false negatives, thereby improving the model's sensitivity in identifying actual positive cases. This novel perspective adds valuable depth to understanding ensemble learning's impact on hepatitis risk prediction, providing a foundation for further advancements in predictive modeling within the healthcare domain.

4. Research Limitations

While the research has yielded valuable insights into the efficacy of Random Forest models enhanced with Bootstrapped Aggregating for hepatitis risk prediction, certain limitations should be acknowledged. The dataset's size and specific attributes may impose constraints on the generalizability of findings to broader populations or varied healthcare contexts. Additionally, the absence of external validation datasets might limit the model's applicability beyond the dataset used for training and testing. Despite these limitations, the study's internal validity remains robust, as the experimental design and methodology align with the research question. The carefully crafted hyperparameter optimization process contributes to the model's reliability, and the consistent improvement in predictive metrics demonstrates the effectiveness of Bootstrapped Aggregating. These limitations prompt a cautious interpretation of the results in broader contexts but do not diminish the relevance and validity of the findings within the specified scope of hepatitis risk prediction.

5. Recommendations for Future Research

For practical implementation, future research could explore integrating the optimized Random Forest model with Bootstrapped Aggregating into real-world healthcare systems for hepatitis risk assessment. Evaluating its performance in a clinical setting and comparing outcomes with traditional diagnostic methods would provide valuable insights into the model's applicability and potential improvements in patient care. Additionally, investigating the impact of diverse datasets from various demographics and regions could enhance the model's generalizability.

Further research could delve into the interpretability of the model's decisions, facilitating its adoption by healthcare professionals. Exploring the scalability and computational efficiency of the proposed model in handling larger datasets would also be crucial for practical implementation. Lastly, investigating ensemble techniques' efficacy in predicting other liver-related conditions could broaden the scope of application and contribute to advancements in predictive modeling within the broader domain of hepatology.

CONCLUSION

This research has yielded significant findings regarding hepatitis risk prediction using a Random Forest model optimized with Bootstrapped Aggregating. With an accuracy reaching 96%, this model demonstrates a substantial improvement compared to the standard Random Forest, which achieved an accuracy of 88%. These findings indicate that applying ensemble techniques, particularly Bootstrapped Aggregating, can enhance the reliability of hepatitis risk predictions. The analysis of the classification reports highlights improvements in precision, recall, and f1-score, with a significant reduction in false negatives. In conclusion, the Random Forest model with Bagging exhibits strong potential for more accurate hepatitis risk assessment. However, it is essential to acknowledge limitations related to dataset size and attribute characteristics and the need for further external validation. This conclusion provides a foundation for developing predictive models in the context of hepatitis risk and offers guidance for future research to enhance the model's usability in clinical settings.

REFERENCES

[1] T. Vos *et al.*, "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *The Lancet*, 2020, doi: 10.1016/s0140-6736(20)30925-9.

[2] C. Gomes, R. J. Wong, and R. G. Gish, "Global Perspective on Hepatitis B Virus Infections in the Era of Effective Vaccines," *Clinics in Liver Disease*, 2019, doi: 10.1016/j.cld.2019.04.001.

[3] M. Jefferies, B. Rauff, H. Rashid, T. M. Lam, and S. Rafiq, "Update on Global Epidemiology of Viral Hepatitis and Preventive Strategies," *World Journal of Clinical Cases*, 2018, doi: 10.12998/wjcc.v6.i13.589.

[4] B. S. Sheena *et al.*, "Global, Regional, and National Burden of Hepatitis B, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *The Lancet Gastroenterology & Hepatology*, 2022, doi: 10.1016/s2468-1253(22)00124-8.

[5] S. G. Sepanlou *et al.*, "The Global, Regional, and National Burden of Cirrhosis by Cause in 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017," *The Lancet Gastroenterology & Hepatology*, 2020, doi: 10.1016/s2468-1253(19)30349-8.

[6] D. Q. Huang *et al.*, "Global Epidemiology of Cirrhosis — Aetiology, Trends and Predictions," *Nature Reviews Gastroenterology & Hepatology*, 2023, doi: 10.1038/s41575-023-00759-2.

[7] O. M. Doyle, N. Leavitt, and J. A. Rigg, "Finding undiagnosed patients with hepatitis C infection: an application of artificial intelligence to patient claims data," *Sci Rep*, vol. 10, no. 1, p. 10521, Jun. 2020, doi: 10.1038/s41598-020-67013-6.

[8] D.-V. Phan, C.-L. Chan, A.-H. A. Li, T.-Y. Chien, and V. L. Nguyen, "Liver Cancer Prediction in a Viral Hepatitis Cohort: A Deep Learning Approach," *International Journal of Cancer*, 2020, doi: 10.1002/ijc.33245.

[9] K. Swetha, A. Kiran, K. Pavanam, E. N. Vijaya Kumari, T. Naresh, and M. J. Baba, "Inflammation of Liver and Hepatitis Disease Prediction using Machine Learning Techniques," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India: IEEE, May 2023, pp. 218–223. doi: 10.1109/ICICCS56967.2023.10142912.

[10] Prasenjit Maity, Arup Kumar Dey, Krishna Prasad Singha, Avijit Kumar Chaudhuri, and Sulekha Das, "An Approach Combining Feature Selection with Machine Learning Techniques for Prediction Reliability and Accuracy in Hepatitis Diagnosis," *IJETMS*, vol. 7, no. 2, pp. 181–194, 2023, doi: 10.46647/ijetms.2023.v07i02.023.

[11] F. R. Albogamy *et al.*, "Decision Support System for Predicting Survivability of Hepatitis Patients," *Frontiers in Public Health*, 2022, doi: 10.3389/fpubh.2022.862497.

[12] B. Khaoula, B. Imene, W. Guenifi, A. Gasmi, and S. Laouamri, "Intelligent Analysis of Some Factors Accompanying Hepatitis B," *Molecular Sciences and Applications*, 2022, doi: 10.37394/232023.2022.2.7.

[13] I. I. Ahmed, D. Y. Mohammed, and K. A. Zidan, "Diagnosis of hepatitis disease using machine learning techniques," *IJEECS*, vol. 26, no. 3, p. 1564, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1564-1572.

[14] M. Ghorbian, "Clinical Usefulness of Machine Learning Approaches as a Non-Invasive Technology in Reducing Hepatitis Disease Mortality," 2023, doi: 10.21203/rs.3.rs-2965115/v1.

[15] S. V. B. -, N. A. -, N. J. -, and M. D. -, "Liver Disease Prediction Using Machine Learning," *International Journal for Multidisciplinary Research*, 2023, doi: 10.36948/ijfmr.2023.v05i03.2955.

[16] 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, p. 481, Apr. 2023, doi: 10.3390/bioengineering10040481.

[17] J. Yang, "Hepatitis C Risk Prediction Based on Adaboost," *Highlights in Science Engineering and Technology*, 2023, doi: 10.54097/hset.v54i.9803.

[18] Η. Δρίτσας and M. Trigka, "Supervised Machine Learning Models for Liver Disease Risk Prediction," *Computers*, 2023, doi: 10.3390/computers12010019.

[19] Z. Farhadi, H. Bevrani, and M. Feizi-Derakhshi, "Improving Random Forest Algorithm by Selecting Appropriate Penalized Method," *Communications in Statistics - Simulation and Computation*, 2022, doi: 10.1080/03610918.2022.2150779.

[20] P. Regier, M. Duggan, A. Myers-Pigg, and N. G. Ward, "Effects of Random Forest Modeling Decisions on Biogeochemical Time Series Predictions," *Limnology and Oceanography Methods*, 2022, doi: 10.1002/lom3.10523.

[21] S. Kanwar, L. K. Awasthi, and V. Shrivastava, "Efficient Random Forest Algorithm for Multi-Objective Optimization in Software Defect Prediction," *Iete Journal of Research*, 2023, doi: 10.1080/03772063.2023.2205377.

[22] R. Susetyoko, E. Purwantini, B. N. Iman, and E. Satriyanto, "An Improved Accuracy of Multiclass Random Forest Classifier With Continuous Attribute Transformation Using Random Percentile Generation," *International Journal on Advanced Science Engineering and Information Technology*, 2023, doi: 10.18517/ijaseit.13.3.18379.

[23] J. Sun and Z. Shen, "Research on Improved Random Forest Algorithm for Highly Unbalanced Data," *Journal of Physics Conference Series*, 2022, doi: 10.1088/1742-6596/2333/1/012007.

[24] S. E. Ibrahim, "Improving Land Use/Cover Classification Accuracy From Random Forest Feature Importance Selection Based on Synergistic Use of Sentinel Data and Digital Elevation Model in Agriculturally Dominated Landscape," *Agriculture*, 2022, doi: 10.3390/agriculture13010098.
[25] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "The Accuracy of Random Forest Performance Can Be Improved by Conducting a Feature Selection With a Balancing Strategy," *Peerj Computer Science*, 2022, doi: 10.7717/peerj-cs.1041.
[26] K. A. Dauda, "Optimal Tuning of Random Survival Forest Hyperparameter With an Application to Liver Disease," *Malaysian Journal of Medical Sciences*, 2022, doi: 10.21315/mjms2022.29.6.7.
[27] Y. Resti, C. Irsan, J. F. Latif, I. Yani, and N. R. Dewi, "A Bootstrap-Aggregating in Random Forest Model for Classification of Corn Plant Diseases and Pests," *Science & Technology Indonesia*, 2023, doi: 10.26554/sti.2023.8.2.288-297.
[28] F. Muhammad *et al.*, "Liver Ailment Prediction Using Random Forest Model," *Computers Materials & Continua*, 2023, doi: 10.32604/cmc.2023.032698.
[29] M. M. Majzoobi, S. Namdar, R. Najafi-Vosough, A. A. Hajilouei, and H. Mahjub, "Prediction of Hepatitis Disease Using Ensemble Learning Methods," *Journal of Preventive Medicine and Hygiene*, p. E424 Pages, Oct. 2022, doi: 10.15167/2421-4248/JPMH2022.63.3.2515.
[30] V. Daniel and R. Ramaraj, "A Novel Modified Long Short Term Memory Architecture for Automatic Liver Disease Prediction From Patient Records," *Concurrency and Computation Practice and Experience*, 2022, doi: 10.1002/cpe.7372.
[31] M. Anisetti, C. A. Ardagna, A. Balestrucci, N. Bena, E. Damiani, and C. Y. Yeun, "On the Robustness of Random Forest Against Untargeted Data Poisoning: An Ensemble-Based Approach," *Ieee Transactions on Sustainable Computing*, 2023, doi: 10.1109/tsusc.2023.3293269.