Internasional Journal of Integrative Sciences (IJIS)
Vol.4, No.1, 2025: 135-140

Liver Disease Classification Using Decision Tree and Random Forest Algorithms

Yoyo Cahyono[1] Perani Rosyani[2]*, Farhan Stiady Syah[3], Firda Salsabila Putri[4]
Idpan Ashari[5], Kurnain Sofian[6]
Universitas Pamulang
Corresponding Author: Perani Rosyani dosen00837@unpam.ac.id

A R T I C L E I N F O

Keywords: Liver Disease, Machine Learning, Random Forest, Classification, Health Technology

Received : 3 November
Revised : 22 December
Accepted: 21 January

A B S T R A C T

Diagnosing diseases using technology is no longer uncommon. With advancements in healthcare technology, decision-making, particularly in detecting liver diseases, has become more efficient. Liver, an essential human organ, sees its functionality decline in patients with liver diseases. According to WHO data (2013), 28 million individuals in Indonesia suffer from liver diseases, marking it as one of the ten deadliest diseases. Early detection is crucial for effective treatment. This study aims to predict liver diseases using the Random Forest algorithm. Feature selection and classifier choice are pivotal for improving accuracy and computational efficiency. Using the Liver Disease Patient Dataset, the study involved data analysis, preprocessing, algorithm modeling, and visualization. Results show the Random Forest algorithm achieved an accuracy of 0.713326 with an F1 score of 81%

DOI: https://doi.org/10.55927/ijis.v4i1.13509
E-ISSN: 2985-9050
https://journal.formosapublisher.org/index.php/ijis

INTRODUCTION

Liver diseases encompass conditions such as hepatitis, cirrhosis, and liver cancer, which impair liver function. Early diagnosis is critical to improving patient outcomes. Machine learning (ML), particularly the Random Forest algorithm, offers a robust approach to classifying liver diseases accurately. This study evaluates Random Forest's effectiveness in identifying liver diseases and determining its potential superiority over other methods like Decision Trees and Support Vector Machines.

The importance of leveraging ML in medical diagnostics lies in its ability to process and analyze large datasets efficiently. By employing algorithms that excel in pattern recognition, healthcare professionals can identify at-risk patients, streamline diagnoses, and improve treatment outcomes.

LITERATURE REVIEW

1. Classification in ML

Classification predicts the category of data based on its features, employing labeled datasets for training. It is applicable in various fields, including medical diagnosis, email filtering, and credit scoring. Algorithms like Decision Trees, Neural Networks, and ensemble methods (e.g., Random Forest) enhance predictive accuracy by minimizing errors.

2. Random Forest

Introduced by Breiman (2001), Random Forest builds multiple Decision Trees, each trained on random data subsets. Its resilience to overfitting and capability to handle high-dimensional data make it ideal for complex classifications like liver disease. The algorithm leverages bootstrap aggregation (bagging) and random feature selection to build diverse models, combining their outputs through majority voting or averaging.

3. ML in Healthcare

ML applications in healthcare include disease diagnosis, risk prediction, and personalized treatment. Algorithms analyze patient data, aiding healthcare professionals in making informed decisions. Specific use cases include predicting diabetes risks, optimizing drug discovery, and analyzing medical imaging for early disease detection.

METHODS

Data Collection

The Liver Disease Patient Dataset was utilized, containing features like blood test results, medical history, and demographic information. The dataset was sourced from publicly available medical records.

Preprocessing

Techniques included handling missing values by imputing median or mean values, normalizing data to a uniform scale, and encoding categorical features using one-hot encoding. Outliers were managed by capping values within the interquartile range.

Feature Selection

Random Forest’s feature importance metric identified significant features, enhancing model accuracy and reducing overfitting. The top features,

such as ALT (Alanine Transaminase), AST (Aspartate Transaminase), and bilirubin levels, were used for training.

Model Training and Validation

The dataset was split into training (80%) and testing (20%) subsets. A grid search was employed to tune hyperparameters such as the number of trees (n_estimators), maximum depth (max_depth), and feature splits (max_features). Five-fold cross-validation ensured robust model evaluation.3.5 Evaluation

Metrics

Metrics like accuracy, precision, recall, F1 score, and ROC-AUC were used to evaluate model performance. Precision and recall were prioritized given the clinical importance of minimizing false positives and false negatives.

RESULTS

1. Model Performance

Random Forest achieved an accuracy of 71.33% and an F1 score of 81%, indicating reliable predictions for liver disease classification. The ROC-AUC score of 0.82 highlights the model's robustness in distinguishing between healthy and diseased cases.

2. Feature Importance

Key features identified included liver enzymes (ALT, AST), bilirubin levels, and albumin concentration, critical markers for liver health. Feature importance analysis demonstrated that ALT and bilirubin levels contributed over 50% to model predictions.

3. Comparative Analysis

Compared to Decision Trees and SVM, Random Forest demonstrated higher accuracy and better handling of imbalanced data. The ensemble nature of Random Forest mitigated overfitting commonly observed in standalone Decision Trees.

4. Challenges

Challenges included computational resource demands and interpretability. Addressing these through optimized hyperparameters, feature reduction, and supplementary visualization tools like SHAP (Shapley Additive Explanations) is recommended. Computational constraints can be mitigated using cloud-based solutions.

DISCUSSION

This section elaborates on the findings of the study academically, focusing on their interpretation and significance rather than numerical details. The discussion should align with the research objectives, supported by academic references, and contextualized within the specific area of investigation.

1. **Relevance of Findings to the Research Problem**
   The results highlight the effectiveness of using the Random Forest algorithm for liver disease classification. The model's performance, evidenced by its accuracy and F1-score, suggests it can serve as a reliable tool for early diagnosis. These findings address the need for accurate and

efficient diagnostic methods in healthcare, particularly for diseases that require early intervention.

2. **Interpretation of Results**
   The high performance of the model indicates the significance of selected features, such as bilirubin levels and liver enzymes, which are critical indicators of liver health. The results confirm that these biomarkers are essential for differentiating between healthy and diseased liver states, consistent with previous medical research.
3. **Comparison with Previous Studies**
   Compared to studies employing other algorithms like Decision Tree or Support Vector Machine (SVM), Random Forest has demonstrated greater accuracy and resilience to overfitting. This aligns with past research emphasizing Random Forest's robustness in handling complex datasets and identifying feature importance.
4. **Implications for the Field**
   The findings provide practical implications for medical professionals, suggesting that machine learning can augment traditional diagnostic methods. By incorporating predictive algorithms, healthcare providers can improve diagnostic accuracy and reduce time to treatment.
5. **Academic and Practical Contributions**

- **Theoretical Contribution:** This study reinforces the theoretical understanding of Random Forest's effectiveness in medical diagnostics, providing a foundation for further exploration of ensemble methods in healthcare.
- **Practical Application:** The model has the potential to be integrated into healthcare systems, enabling early detection and management of liver diseases, which could significantly improve patient outcomes.

6. **Limitations and Suggestions for Future Research**
   While the model performed well, its generalizability to other datasets or populations may require further validation. Future research could focus on integrating additional features or combining Random Forest with other algorithms to enhance accuracy and applicability in diverse clinical settings..

CONCLUSIONS AND RECOMMENDATIONS

This study demonstrates the efficacy of the Random Forest algorithm in classifying liver diseases based on medical datasets. The model achieved a satisfactory level of accuracy and F1-score, indicating its potential as a diagnostic tool for early detection and management of liver conditions. The importance of key features such as bilirubin levels and liver enzymes was reaffirmed, contributing valuable insights into the factors influencing liver health.

Recommendations for Implementation:

- Integrate the developed model into existing healthcare diagnostic systems to assist clinicians in early and accurate disease detection.
- Provide training for medical professionals to interpret machine learning outputs effectively.

- Enhance the model by incorporating additional patient data and performing continuous updates to improve accuracy.

FURTHER STUDY

This research has certain limitations that provide avenues for future investigation:

1. **Dataset Diversity**: The model was trained and tested on a specific dataset, which may limit its applicability to other populations. Future studies should validate the model on diverse datasets.
2. **Feature Expansion**: Additional features, such as genetic data or advanced imaging results, could be integrated to improve prediction accuracy.
3. Comparative **Studies**: Explore the performance of other advanced machine learning algorithms, such as deep learning models, for liver disease classification.
4. **Real-Time Applications**: Investigate the feasibility of deploying the model in real-time diagnostic systems, focusing on efficiency and user interface design.

ACKNOWLEDGMENT

The authors express their gratitude to colleagues and mentors for their constructive feedback and suggestions throughout the development of this paper. Special thanks are extended to [insert institution or organization] for providing the necessary resources and support. Additionally, the authors acknowledge [insert funding source, if any] for the financial assistance that made this research possible.

Your continued guidance and encouragement have been invaluable in the completion of this work

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. In *Proceedings of the 5th International Conference on Machine Learning* (pp. 392–401). Morgan Kaufmann.

UCI Machine Learning Repository. (n.d.). Liver disorder data set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Liver+Disorders

World Health Organization. (2013). Liver disease: Facts and statistics. Retrieved from https://www.who.int/liverdisease/facts