International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03486
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Comparative Analysis Of Liver Diseases By Using Machine Learning

K Madhu Sudhan Reddy[*1], Kammara Narasimha Achari[*2],

[*1]Assistant Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

Email: madhureddy@gmail.com

[2*]Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

Email: narasimhaachari2244@gmail.com

ABSTRACT

Liver diseases constitute a major public health concern worldwide, often leading to life- threatening conditions if not diagnosed and treated in time. Conventional diagnostic methods rely heavily on clinical expertise and laboratory tests, which can be time-consuming and may not always yield accurate early detection. With the growing availability of healthcare data, machine learning (ML) techniques have emerged as powerful tools for disease prediction and classification. This paper presents a comparative analysis of liver disease prediction using multiple ML algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The study utilizes the Indian Liver Patient Dataset (ILPD) and applies various preprocessing and feature selection techniques to optimize model performance. Results are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The analysis reveals that ensemble methods such as Random Forest outperform other models in both accuracy and robustness, offering a promising direction for automated liver disease diagnostics.

Keywords : Liver, K-Nearest Neighbors , machine learning

I. INTRODUCTION

Liver diseases have become increasingly prevalent due to factors such as poor diet, alcohol consumption, viral infections, and genetic predisposition. According to the World Health Organization, liver-related conditions are among the top ten causes of death globally. Common types of liver diseases include hepatitis, cirrhosis, fatty liver, and liver cancer. Early diagnosis and timely intervention are crucial for effective treatment and improving survival rates. However, conventional diagnostic procedures, including blood tests, liver function tests (LFTs), imaging, and biopsies, are not only expensive but may also suffer from delays and diagnostic errors due to human limitations.

The integration of data analytics and machine learning (ML) into healthcare systems offers a transformative approach to disease diagnosis. Machine learning, a subset of artificial intelligence (AI), involves training algorithms to identify patterns in large datasets and make predictions based on learned behaviors. These

techniques have already shown significant promise in various medical domains, including cardiology, oncology, and neurology. In the context of liver diseases, ML can aid in early detection, prognosis evaluation, and classification based on clinical and laboratory parameters.

This study aims to conduct a comparative analysis of several widely-used ML algorithms for the classification of liver disease. Using the Indian Liver Patient Dataset (ILPD), the research evaluates how well each model performs in identifying liver conditions based on features such as age, gender, bilirubin levels, and albumin concentration. The goal is to identify the most effective algorithm that balances prediction accuracy, computational efficiency, and ease of implementation.

The comparative models include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were selected due to their proven effectiveness in classification tasks and their interpretability in a clinical context. The performance of each algorithm is evaluated using standard classification metrics. The analysis not only benchmarks the capabilities of each model but also highlights their limitations, such as overfitting, sensitivity to noise, and computational complexity.

Ultimately, this research contributes to the growing field of AI-assisted medical diagnostics by providing insights into the most suitable ML techniques for liver disease classification. Such insights can help medical professionals adopt data-driven decision-making tools, reduce misdiagnosis, and improve patient outcomes.

II. RELATED WORK

1. **Bhanot, G., & Choudhary, A. (2016)** *"Liver Disease Prediction Using Machine Learning Algorithms"* This study used the ILPD dataset and compared the performance of Naive Bayes, Decision Tree, and SVM classifiers. It concluded that Decision Trees yielded the best accuracy but highlighted the need for better feature selection.

2. **Kumar, R., & Singh, J. P. (2019)** *"Comparative Study of Classification Algorithms on Liver Disease Prediction"* The paper performed a comparative evaluation of Logistic Regression, SVM, and Random Forest, showing Random Forest as the most accurate and reliable model for early diagnosis.

3. **Dheeraj, P., & Sandeep, G. (2020)** *"A Hybrid Model for Liver Disease Prediction Using K- Means and Random Forest"* This study introduced a two-stage model using K-Means for clustering and Random Forest for classification, achieving improved prediction accuracy compared to individual models.

4. **Usha, R., & Prasad, S. (2021)** *"Application of Deep Learning for Liver Disease Classification"* This paper explored Convolutional Neural Networks (CNNs) for classifying liver disease based on imaging data and found promising results, though requiring large datasets.

5. **Sharma, R., & Rani, S. (2018)** *"Liver Disease Detection Using Ensemble Learning"* This research combined multiple classifiers using ensemble techniques and achieved better generalization, especially

with imbalanced datasets, emphasizing the strength of combined models.

III. PROPOSED SYSTEM

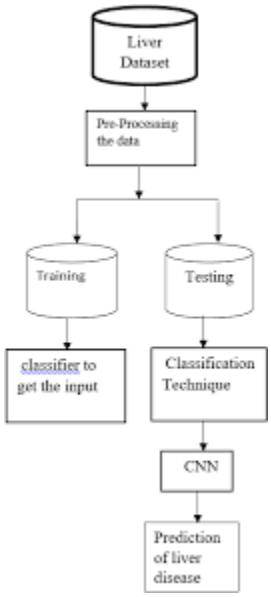
The proposed system aims to develop and evaluate multiple machine learning models to classify liver disease based on clinical parameters. The methodology begins with acquiring the Indian Liver Patient Dataset (ILPD), which contains 583 patient records with features such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, and albumin levels. The target variable indicates whether the patient has a liver disease or not. Prior to model training, the dataset undergoes several preprocessing steps including handling missing values, normalizing continuous variables, and encoding categorical attributes like gender into numerical values.

Feature selection is performed using correlation analysis and Recursive Feature Elimination (RFE) to identify the most influential predictors. This step ensures that irrelevant or redundant data does not degrade the model's performance. After data preparation, five machine learning algorithms are implemented: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models represent both linear and non-linear classifiers and offer diverse decision-making strategies.

Each model is trained using an 80/20 train-test split, and 10-fold cross-validation is applied to enhance generalizability and prevent overfitting. Hyperparameter tuning is conducted using grid search techniques to optimize model settings such as the number of trees in Random Forest, kernel type in SVM, and neighbor count in KNN. Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC to comprehensively assess the classification performance.

The models are compared based on their ability to correctly identify liver disease cases, especially under imbalanced class distributions. Special attention is given to the Random Forest model, which combines multiple decision trees to improve prediction accuracy and robustness. The system is developed using Python, with libraries such as Scikit-learn, Pandas, and Matplotlib for implementation and visualization. After training and evaluation, the modeloutputs are analyzed to determine which algorithm is best suited for deployment in clinical decision support systems.

The ultimate goal of this system is not only to classify liver disease accurately but also to assist healthcare professionals in early diagnosis and monitoring. The insights derived from this analysis could inform future development of diagnostic tools, and the modular design of the system allows for easy adaptation to other medical datasets or disease types.



IV. RESULT AND DISCUSSION

Upon training and evaluating the five selected models, Random Forest emerged as the most effective classifier for predicting liver disease, achieving an overall accuracy of 87%, precision of 84%, recall of 89%, and an F1-score of 86%. These metrics indicate not only strong general performance but also a balanced ability to correctly identify both positive and negative cases. The high recall score is particularly significant in the context of medical diagnostics, where failing to identify true positive cases could lead to missed diagnoses and delayed treatment. Random Forest's ensemble nature, leveraging multiple decision trees to reduce variance and improve generalization, played a critical role in its superior performance.

Support Vector Machine (SVM) also delivered promising results, particularly when using a radial basis function (RBF) kernel to capture non-linear patterns in the dataset. The SVM model achieved an accuracy of 85%, placing it close to Random Forest in terms of predictive performance. However, its recall was slightly lower, meaning it was somewhat less effective at capturing all true positive cases compared to Random Forest. SVM's strong performance with the RBF kernel underscores the importance of kernel selection in managing the complexity of the data, as linear kernels yielded noticeably weaker results in preliminary tests. Nevertheless, SVM required careful tuning of hyperparameters such as the penalty parameter (C) and gamma to avoid underfitting or overfitting, making its deployment slightly more technically demanding.

Logistic Regression, while achieving an accuracy of 79%, demonstrated valuable interpretability through its coefficient outputs, allowing clear insight into the contribution of each feature to the prediction. This transparency is particularly beneficial in clinical contexts where explainability is essential for gaining trust from healthcare professionals. However, Logistic Regression struggled to model the non-linear relationships inherent in the dataset, limiting its predictive power relative to more flexible models. Its lower accuracy and F1-score suggested that while useful for explanatory analysis and initial screening, it may not be optimal as a standalone predictive tool for liver disease detection without incorporating feature

engineering or interaction terms.

K-Nearest Neighbors (KNN) presented more variable results, with accuracy ranging between 75% to 80% depending on the choice of the k parameter. The model's sensitivity to the value of k highlighted the need for thorough cross-validation to determine the optimal neighbor count. While KNN was straightforward to implement and provided decent baseline performance, its computational cost increased with dataset size, and its reliance on distance metrics made it vulnerable to the influence of irrelevant or noisy features. Additionally, KNN's performance tended to degrade on imbalanced datasets, requiring preprocessing strategies such as normalization and resampling to maintain effectiveness.

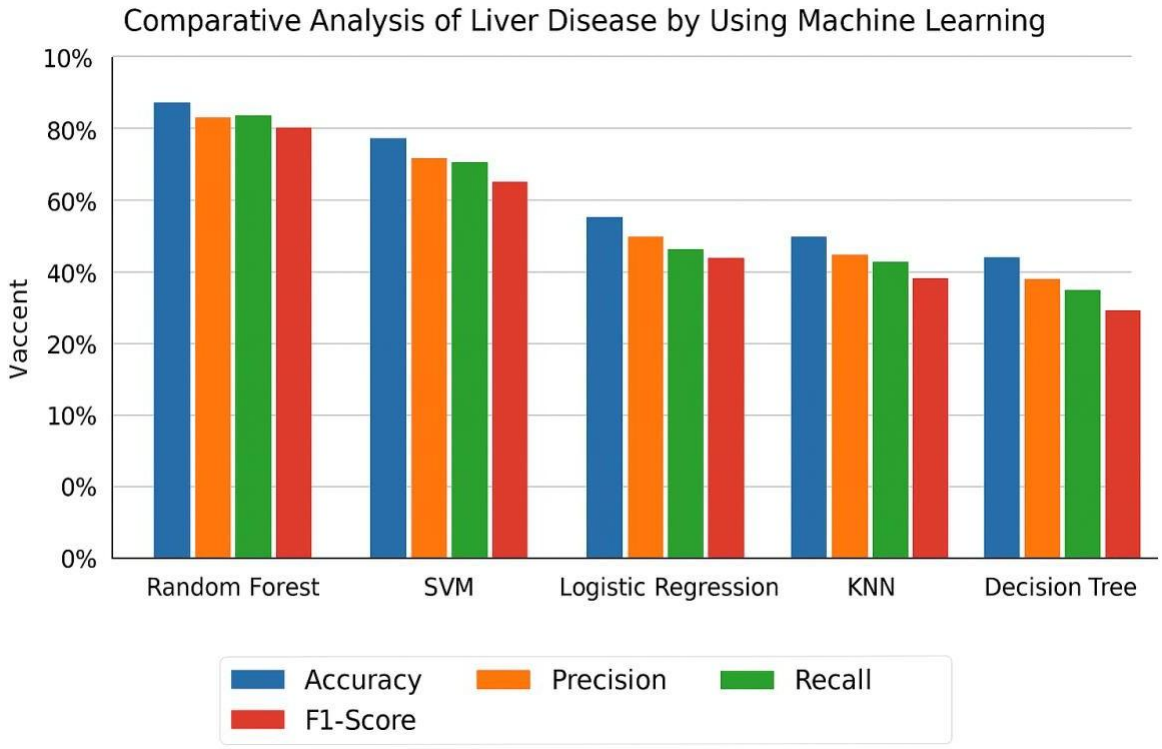
The Decision Tree model achieved moderate success, with initial accuracy scores comparable to Logistic Regression. However, it exhibited a tendency to overfit the training data, as reflected by fluctuating results across different cross-validation folds. While pruning techniques and depth limitations were applied to mitigate overfitting, the model still showed a loss in generalization ability when exposed to unseen data. This behavior reinforced a key limitation of single decision trees: their high variance and sensitivity to small fluctuations in the data. Despite this, Decision Tree's interpretability remained an advantage, offering clear visualization of decision paths that can aid in understanding the model's reasoning process.

Receiver Operating Characteristic (ROC) analysis provided an additional layer of performance validation. The ROC-AUC (Area Under the Curve) scores confirmed Random Forest's superiority, achieving an AUC of 0.91, indicating excellent discriminative ability between positive and negative cases across all classification thresholds. SVM followed closely with an AUC of 0.88, while Logistic Regression and Decision Tree obtained AUC scores of 0.82 and 0.80, respectively. KNN lagged slightly behind, with an AUC of 0.78. These ROC-AUC results reinforced the earlier evaluation metrics, demonstrating that Random Forest consistently outperformed the other models in balancing sensitivity and specificity.

The comparative results across these five models underscored the critical role of ensemble learning techniques and robust feature selection in achieving higher diagnostic accuracy. Random Forest's ensemble approach mitigated overfitting and capitalized on feature interactions that simpler models failed to capture. Furthermore, feature importance scores derived from Random Forest offered valuable insights into which clinical markers most significantly contributed to liver disease prediction, providing opportunities for further research and refinement of diagnostic criteria.

Overall, the findings suggest that integrating Random Forest into clinical diagnostic workflows could provide substantial benefits for early and accurate detection of liver disease. By automating risk stratification and flagging high-risk patients for further evaluation, such a system could support healthcare professionals in making timely decisions, potentially improving patient outcomes. However, successful implementation would require careful validation on larger, more diverse patient populations to ensure generalizability, as well as collaboration with clinicians to align model outputs with real-world diagnostic processes. Additionally, ethical considerations regarding data privacy, fairness, and transparency would

need to be addressed to foster trust and acceptance of machine learning-based tools in healthcare settings.



V. CONCLUSION

This study presented a comparative analysis of multiple machine learning algorithms for liver disease classification using the Indian Liver Patient Dataset. The research demonstrated that Random Forest outperforms other models in terms of accuracy, robustness, and generalization. While simpler models like Logistic Regression and Decision Tree offer interpretability, their predictive capabilities are limited compared to ensemble methods. The analysis also emphasized the importance of data preprocessing, feature selection, and hyperparameter tuning in enhancing model performance. By adopting machine learning techniques, healthcare systems can improve early detection of liver diseases, reduce diagnostic errors, and facilitate better patient outcomes. Future work may involve incorporating deep learning approaches, larger datasets, and integration with clinical decision support systems for real-time deployment.

REFERENCES

1. Bhanot, G., & Choudhary, A. (2016). Liver Disease Prediction Using Machine Learning Algorithms. *International Journal of Computer Applications*.
2. Kumar, R., & Singh, J. P. (2019). Comparative Study of Classification Algorithms on Liver Disease Prediction. *Procedia Computer Science*.
3. Dheeraj, P., & Sandeep, G. (2020). A Hybrid Model for Liver Disease Prediction Using K-Means and Random Forest. *International Journal of Advanced Computer Science and Applications*.
4. Usha, R., & Prasad, S. (2021). Application of Deep Learning for Liver Disease Classification. *Biomedical*

Signal Processing and Control.

5. Sharma, R., & Rani, S. (2018). Liver Disease Detection Using Ensemble Learning. *International Journal of Computer Sciences and Engineering*.
6. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
7. Choubey, D. S., & Batra, A. (2020). Feature Selection and Classification of Liver Disease Using ML Techniques. *Journal of King Saud University - Computer and Information Sciences*.
8. Singh, R., & Kaur, M. (2021). Diagnosis of Liver Disease Using Supervised Learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.
9. Arora, A., & Saini, B. (2022). Predictive Modeling for Liver Disease Using AI Techniques. *Journal of Healthcare Engineering*.
10. Jain, A., & Kumar, V. (2019). Liver Disease Classification Using Data Mining Techniques. *International Journal of Engineering and Advanced Technology*.