

AI-Based Early Disease Prediction System: A Comparative and Explainable Machine Learning Approach for Chronic Kidney and Liver Diseases

Jata Shankar, Aniket Dubey, Vijay Kage, Sahil Dhane

Students, Department of Master of Computer Application.

D. Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune, Maharashtra, India.

Guide: Ms. Yogeshwari Yawalkar

ABSTRACT

Early prediction of Chronic Kidney Disease (CKD) and liver disease can support timely medical assessment and improve disease management. However, identifying these conditions through clinical and laboratory data can be challenging due to variations in patient characteristics and the complexity of medical attributes. This study proposes an AI-Based Early Disease Prediction System using machine learning techniques to analyse CKD and liver disease datasets. Five algorithms—Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM—will be evaluated to compare their classification performance for both diseases. Data preprocessing techniques will be applied to handle missing values, encode categorical features, and prepare the datasets for model training and testing. Performance will be assessed using accuracy, precision, recall, F1-score, and ROC-AUC, where applicable. In addition, SHapley Additive exPlanations (SHAP) will be used to interpret model predictions and identify features contributing to the output. The proposed framework aims to provide a comparative and interpretable approach to preliminary disease-risk prediction. As a research prototype, it may support healthcare research and early risk assessment but is not intended to replace professional diagnosis. Further testing and independent clinical validation will be necessary before practical clinical use.

Keywords: Artificial Intelligence, Chronic Kidney Disease, Liver Disease, Machine Learning, SHAP, Explainable AI.

INTRODUCTION

Chronic Kidney Disease (CKD) and liver disease are important health conditions that can affect normal body functioning and may lead to serious complications when not identified in time. Their early identification can help healthcare professionals assess a patient's condition and make appropriate decisions about further testing and treatment. However, recognizing disease-related patterns from clinical and laboratory measurements can be challenging because patient data may contain missing values, variations, and relationships that are difficult to identify through manual analysis alone.

Machine learning offers a data-driven approach to analysing medical attributes and classifying patients according to patterns associated with particular diseases. Different algorithms may produce different prediction results depending on the dataset and the characteristics of the input features. Therefore, evaluating multiple algorithms is important for understanding their relative performance rather than relying on a single model. In this study, Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM are proposed for two separate prediction tasks: CKD prediction and liver disease prediction.

Although predictive models can assist in identifying patterns in healthcare data, their outputs may be difficult to interpret without additional explanation. To address this concern, the proposed system incorporates SHapley Additive exPlanations (SHAP), which can indicate how individual features contribute to a model's predictions. This provides a way to examine both overall feature importance and explanations for individual predictions, while recognizing that feature contributions do not establish medical causation.

This research proposes an AI-based framework that combines disease-specific machine learning models with explainability techniques. The models will be evaluated using suitable classification metrics to compare their performance on the two datasets. The project aims to support preliminary disease-risk assessment and contribute to research in interpretable healthcare applications, aligned with Sustainable Development Goal 3: Good Health and Well-being. The proposed system is intended as a research prototype and not as a replacement for professional medical diagnosis.

PROBLEM STATEMENT

Predicting Chronic Kidney Disease (CKD) and liver disease from clinical data is challenging because relevant patient attributes may contain missing values and complex patterns. A single machine learning algorithm may not perform equally well on both disease datasets, making model comparison important. Additionally, prediction results alone may not clearly explain which patient attributes influence the model's output.

Therefore, this research proposes separate prediction models for CKD and liver disease using five machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. SHAP will be incorporated to explain feature contributions, while suitable evaluation metrics will be used to compare model performance and support interpretable preliminary disease-risk assessment.

OBJECTIVES

1. To develop separate machine learning models for predicting Chronic Kidney Disease (CKD) and liver disease using clinical datasets.
2. To compare Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM using suitable classification metrics.
3. To apply SHAP to identify the features contributing to model predictions and improve interpretability.
4. To develop a prototype that presents prediction results and explanations for preliminary disease-risk assessment.

LITERATURE REVIEW

Previous studies have explored machine learning for Chronic Kidney Disease (CKD) prediction and artificial intelligence for liver disease assessment. Research on CKD has examined classification algorithms and explainable AI to identify important clinical features and improve the interpretation of predictions [1, 2]. Other studies have reviewed machine learning applications in CKD progression, highlighting the potential of these methods as well as challenges related to data quality and model validation [3]. Research on liver disease has also investigated AI-based approaches for analysing clinical data and supporting disease assessment [4]. These studies provide a foundation for applying machine learning to both diseases. However, comparing multiple algorithms separately on CKD and liver disease datasets, while using SHAP to explain model predictions, can provide a focused approach to evaluating performance and interpretability within the proposed system.

RESEARCH GAP

Previous studies have demonstrated the potential of machine learning for Chronic Kidney Disease (CKD) prediction, CKD progression assessment, and AI-assisted liver disease diagnosis. However, the reviewed studies differ in their datasets, prediction objectives, and evaluation methods.

Based on the reviewed literature, the present study identifies the following areas for investigation:

1. **Comparative evaluation:** Further comparative evaluation of Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM across separate CKD and liver disease prediction tasks is needed within the proposed project framework.
2. **Explainability:** Integrating SHAP with the selected models can help examine the contribution of clinical and laboratory features to individual predictions.
3. **Disease-specific evaluation:** Since the two datasets may differ in features, class distribution, and target definitions, each disease-prediction task requires separate performance evaluation.
4. **Practical implementation:** There is a need to investigate how comparative model results and feature explanations can be presented through a prototype while clearly communicating the limitations of predictions based on public datasets.

To address these areas, this study proposes separate CKD and liver disease classification models using five machine learning algorithms, evaluates their performance using suitable classification metrics, and incorporates SHAP-based explanations. The identified gaps will be assessed through the proposed experiments rather than assumed to be resolved in advance.

METHODOLOGY

1. Data Sources

The proposed system uses two publicly available datasets:

- CKD Dataset: Approximately 400 records containing clinical attributes related to Chronic Kidney Disease.
- Indian Liver Patient Dataset (ILPD): Approximately 583 records containing patient attributes relevant to liver disease prediction.

2. Preprocessing Techniques

The following preprocessing steps will be applied:

- Handling missing values and duplicate records.
- Encoding categorical attributes into numerical values.
- Applying feature scaling where required.
- Splitting each dataset into training and testing sets using an 80:20 ratio.
- Fitting preprocessing steps on training data to prevent data leakage.

3. Machine Learning Algorithms

Five supervised learning algorithms will be evaluated separately for CKD and liver disease prediction:

1. Logistic Regression: Provides a baseline classification model.
2. Decision Tree: Learns decision rules from patient attributes.
3. Random Forest: Combines multiple decision trees to improve classification.
4. XGBoost: Uses gradient boosting to learn complex patterns in data.
5. LightGBM: Provides an efficient gradient-boosting approach for classification.

4. Explainable Artificial Intelligence

SHapley Additive exPlanations (SHAP) will be used to understand how individual features contribute to model predictions. Global explanations will summarize influential features, while local explanations will describe contributions to individual predictions.

5. Evaluation Metrics

The models will be evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC, where applicable
- Confusion matrix

These metrics will be calculated separately for the two disease-prediction tasks to support a fair comparison of model performance.

6. Tools and Technologies

Python will be used with Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost, LightGBM, and SHAP. Google Colab is proposed as the development environment, with Streamlit considered for presenting the system through a user interface.

RESULTS AND DISCUSSION

The proposed system will compare Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM for Chronic Kidney Disease (CKD) and liver disease prediction. Each algorithm will be evaluated separately using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, where applicable. The comparison will help identify differences in classification performance across the two datasets.

SHAP will be used to examine the contribution of individual features to model predictions. These explanations may help clarify which clinical attributes influence the models' outputs and improve the interpretability of the proposed system. However, the importance of a feature in a model does not necessarily indicate a medical cause of disease.

The study may face limitations related to dataset size, missing values, class imbalance, and differences in patient characteristics. Moreover, results obtained from public datasets may not represent performance in other clinical settings. Therefore, conclusions about model effectiveness will depend on the completed experiments, and independent validation will be necessary before considering practical clinical use.

CONCLUSION

The proposed AI-Based Early Disease Prediction System aims to support the preliminary prediction of Chronic Kidney Disease (CKD) and liver disease using machine learning techniques. By comparing Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, the study will examine the performance of different algorithms on disease-specific datasets. The integration of SHAP will provide explanations of feature contributions, making model predictions easier to interpret.

The system may contribute to healthcare research by combining predictive modelling with explainable AI. However, its reliability will depend on data quality, experimental results, and validation on independent datasets. Future work should focus on testing the models with larger and more diverse patient data, improving generalizability, and assessing their suitability for clinical settings. The proposed framework is intended to assist preliminary risk assessment and research, not replace medical professionals or clinical diagnosis.

EXPECTED OUTCOME AND FUTURE SCOPE

The study is expected to produce a comparative evaluation of Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM for separate CKD and liver disease prediction tasks. SHAP will be used to examine the contribution of input features. The actual performance of each model will be reported after experiments are completed.

Future work may include external validation, larger and more diverse datasets, improved handling of class imbalance, and collaboration with healthcare professionals to assess the system's practical relevance.

REFERENCES

1. Dharmarathne, G. T., et al. (2024). On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence. *Intelligent Systems with Applications*, 22, 200397. <https://doi.org/10.1016/j.iswa.2024.200397>
2. Isaza-Ruget, M. A., Yomayusa, N., González, C. A., et al. (2024). Predicting chronic kidney disease progression with artificial intelligence. *BMC Nephrology*, 25, 148. <https://doi.org/10.1186/s12882-024-03545-7>
3. Khalid, F., Alsadoun, L., Khilji, F., et al. (2024). Predicting the progression of chronic kidney disease: A systematic review of artificial intelligence and machine learning approaches. *Cureus*, 16(5), e60145. <https://doi.org/10.7759/cureus.60145>
4. Decharatanachart, P., Chaiteerakij, R., Tiyyarattanachai, T., & Treeprasertsuk, S. (2021). Application of artificial intelligence in chronic liver diseases: A systematic review and meta-analysis. *BMC Gastroenterology*, 21, 10. <https://doi.org/10.1186/s12876-020-01585-5>
5. Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal*, 6, 100169. <https://doi.org/10.1016/j.dajour.2023.100169>
6. Jawad, K. M. T., Verma, A., Amsaad, F., & Ashraf, L. (2024). AI-driven predictive analytics approach for early prognosis of chronic kidney disease using ensemble learning and explainable AI. *arXiv preprint*. <https://arxiv.org/abs/2406.06728>
7. Kannan, S. K. N., Aseervatham, J., Moholkar, K., Palanimuthu, M., Marappan, S., Muthusamy, N., Sathar, B., & Sengan, S. (2024). A model for predicting chronic kidney diseases based on medical data using reinforcement learning. *SN Computer Science*, 5, 353. <https://doi.org/10.1007/s42979-024-02665-z>
8. Kumar, S., & Rani, P. (2024). An AI-based liver disease prediction model based on Pearson correlation feature selection method. *Biomedical and Pharmacology Journal*, 17(4), 2187–2202. <https://doi.org/10.13005/bpj/3016>
9. Islam, M. R., Kowser, S., Chowdhury, A. I., et al. (2025). Early prediction of liver disease using ensemble machine learning models and explainable AI techniques. *Biomedical Materials & Devices*, 4(3), 3601–3638. <https://doi.org/10.1007/s44174-025-00410-5>
10. Topcu, A. E., Elbaşı, E., & Alzoubi, Y. I. (2024). Machine learning-based analysis and prediction of liver cirrhosis. In *Proceedings of the International Conference on Telecommunications and Signal Processing*. <https://doi.org/10.1109/TSP63128.2024.10605929>
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://arxiv.org/abs/1705.07874>
12. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
13. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154. <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
15. Ganie, S. M., Dutta Pramanik, P. K., Mallik, S., & Zhao, Z. (2023). Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PLOS ONE*, 18(12), e0295234. <https://doi.org/10.1371/journal.pone.0295234>
16. Nam, D., Chapiro, J., Paradis, V., Seraphin, T. P., Kather, J. N., & Kitano, M. (2022). Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. *JHEP Reports*, 4(4), 100443. <https://doi.org/10.1016/j.jhepr.2022.100443>