IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

Predicting Diabetes, Heart, Liver, and Kidney Disease Using Gradient Boosting and Explainability Frameworks

Ishant[1], Hemant Kumar[2], Rajendra Singh[3]
[1]Student, [2]Associate Professor, [3]Dean,
Department of Computer Science and Engineering,
Raffles University, Neemrana, Rajasthan, India.
[1] yishant888@gmail.com [2] hemantkumar@rafflesuniversity.edu.in [3] rajendra.singh@rafflesuniversity.edu.in

Abstract: *This paper presents a benchmark study for chronic disease prediction covering diabetes, heart disease, liver disease, and chronic kidney disease using classical machine learning models. Logistic Regression, Random Forest, and XGBoost classifiers are systematically evaluated under identical experimental conditions using leakage-safe preprocessing pipelines, SMOTE-based class balancing within cross-validation folds, 5-fold stratified cross-validation, GridSearchCV hyperparameter tuning, and SHAP-based post-hoc explainability analysis. Results show that gradient boosting and ensemble methods achieve competitive or superior performance compared to linear baselines across most disease datasets, with perfect classification achieved on chronic kidney disease. The framework is intended for educational and research purposes and does not serve as a clinical diagnostic system.*

Keywords: chronic disease prediction, gradient boosting, XGBoost, Random Forest, SMOTE, SHAP explainability, benchmark study, clinical machine learning, cross-validation

I. INTRODUCTION

Chronic diseases including diabetes, heart disease, liver disease, and chronic kidney disease collectively represent a major global health burden. According to the World Health Organization, non-communicable diseases account for the majority of deaths worldwide, with many cases remaining undetected until advanced stages. Early and accurate prediction of disease risk from structured clinical data can support timely intervention and reduce long-term health complications.

Machine learning has emerged as a powerful tool for clinical risk prediction, particularly when applied to structured tabular datasets containing patient measurements and laboratory results. Classical algorithms such as Logistic Regression, Random Forest, and XGBoost have demonstrated strong performance across a wide range of medical prediction tasks. However, most published studies evaluate models on a single disease dataset or use inconsistent preprocessing protocols, making cross-study comparisons unreliable.

This paper addresses these limitations by conducting a systematic benchmark study across four disease datasets using a unified experimental framework. Every dataset undergoes the same preprocessing pipeline, the same model training procedure, and the same evaluation protocol. This ensures that observed performance differences reflect genuine dataset and model characteristics rather than methodological inconsistencies.

A further contribution of this work is the integration of SHAP-based explainability, which provides feature-level interpretations of model predictions. Explainability is increasingly recognized as a critical requirement for machine learning systems deployed in healthcare contexts, as it allows clinicians and researchers to verify that model decisions align with clinical knowledge.

The remainder of this paper is organized as follows. Section II reviews related literature. Section III describes the datasets, preprocessing pipeline, model descriptions, and experimental setup. Section IV presents results and discussion. Section

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

V covers explainability analysis. Section VI discusses limitations. Section VII outlines future work. Section VIII concludes the paper.

II. LITERATURE REVIEW

A. Machine Learning in Clinical Prediction

The application of machine learning to clinical prediction tasks has grown substantially over the past two decades. Early work demonstrated that decision tree-based models could match or exceed logistic regression on structured medical datasets. Subsequent studies introduced ensemble methods that further improved robustness by combining multiple weak learners into a single strong predictor.

Nusinovici et al. (2020) conducted a large-scale comparison of logistic regression and machine learning models for predicting major chronic diseases and found that logistic regression remained competitive with more complex models when features were well-engineered. Fernández-Delgado et al. (2014) evaluated 179 classifiers across 121 datasets and found that Random Forest achieved the highest average accuracy, establishing it as a strong general-purpose baseline.

B. Diabetes Prediction

The Pima Indians Diabetes Database has been widely used as a benchmark for diabetes prediction research. Reported accuracy values for classical models on this dataset range from approximately 0.70 to 0.80. Studies have shown that glucose concentration, BMI, and age are consistently among the most predictive features. Class imbalance between diabetic and non-diabetic cases is a known challenge, and SMOTE-based oversampling has been shown to improve recall for the minority diabetic class.

C. Heart Disease Prediction

The Cleveland Heart Disease dataset from the UCI Machine Learning Repository is the most commonly used benchmark for cardiac risk prediction. Ensemble models and gradient boosting methods have achieved AUC values above 0.93 on this dataset. Logistic Regression has also shown strong performance due to the relatively linear separability of the feature space in this dataset.

D. Liver Disease Prediction

The Indian Liver Patient Dataset presents classification challenges due to significant class imbalance, with liver disease patients outnumbering healthy controls. Studies have reported that precision and recall trade-offs are particularly important in this setting. XGBoost and Random Forest have shown stronger recall performance compared to Logistic Regression on this dataset.

E. Chronic Kidney Disease Prediction

The Chronic Kidney Disease dataset from the UCI Machine Learning Repository contains 24 features including both numerical and categorical attributes. Tree-based models have consistently achieved near-perfect classification on this dataset, likely due to the high discriminative power of features such as hemoglobin and serum creatinine.

F. Explainability

Lundberg and Lee (2017) introduced SHAP values as a unified framework for interpreting machine learning model predictions based on Shapley values from cooperative game theory. SHAP provides both local explanations for individual predictions and global explanations summarizing feature importance across the dataset.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

III. METHODOLOGY

A. Dataset Summary

Four publicly available clinical datasets from the UCI Machine Learning Repository are used in this study. Table I summarizes the key characteristics of each dataset.

Dataset	Samples	Features	Target
Diabetes	768	8	Outcome
Heart Disease	297	13	target
Liver Disease	583	10	Selector
Chronic Kidney Disease	400	24	class

The Diabetes dataset contains clinical measurements including glucose concentration, BMI, blood pressure, insulin level, skin thickness, diabetes pedigree function, number of pregnancies, and age. The Heart Disease dataset contains features such as chest pain type, resting blood pressure, cholesterol, maximum heart rate, and exercise-induced angina. The Liver Disease dataset contains features including total bilirubin, direct bilirubin, alkaline phosphatase, and albumin levels. The Chronic Kidney Disease dataset contains both numerical features such as hemoglobin and serum creatinine, and categorical features such as presence of hypertension and diabetes mellitus.

B. Preprocessing Pipeline

All datasets undergo a standardized preprocessing workflow designed to prevent data leakage and ensure fair model comparison. The pipeline consists of the following sequential steps.

Missing values are identified and imputed using median imputation. Median imputation is preferred over mean imputation because it is robust to outliers commonly present in clinical measurements. Categorical features are encoded using label encoding to convert string labels into integer representations compatible with all three model families.

Feature standardization is applied using StandardScaler, which transforms each numerical feature to zero mean and unit variance according to:

$$z = (x - \mu) / \sigma$$

where μ is the mean of the feature computed on the training set and σ is the corresponding standard deviation. The scaler is fit exclusively on training data and subsequently applied to both validation and test data without refitting, preventing any information from the validation or test distributions from influencing the scaling parameters.

C. Model Descriptions

Three model families are selected for evaluation based on their widespread use in clinical prediction literature and their representation of distinct learning paradigms.

Logistic Regression is a linear probabilistic classifier that estimates the probability of class membership using the sigmoid activation function:

$$P(y=1|x) = 1 / (1 + e^{(-z)})$$

where $z = w^T x + b$ is a linear combination of input features weighted by learned parameters w and bias b. The decision boundary is the hyperplane where the predicted probability equals the classification threshold. Logistic Regression serves as an interpretable linear baseline against which ensemble methods are compared.

Random Forest is a bagging ensemble of decision trees, each trained on a bootstrap sample of the training data with random feature subsampling at each node split. Node splitting selects the feature and threshold that minimizes Gini impurity:

$$Gini = 1 - \Sigma(p\_i^2)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

where p_i is the proportion of samples belonging to class i within a node. Final class prediction is determined by majority vote across all trees in the ensemble. Random Forest reduces variance compared to individual decision trees while maintaining strong non-linear modeling capability.

XGBoost is a regularized gradient boosting framework that builds an additive ensemble of decision trees sequentially, where each new tree is trained to correct the residual errors of the current ensemble. The training objective minimizes:

$$\text{Obj} = \Sigma\, l(y_i, \hat{y}_i) + \Sigma\, \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is the per-sample loss function and the regularization term $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes model complexity through the number of leaves T and the magnitude of leaf weights w. The regularization terms γ and λ control the degree of complexity penalization, improving generalization on tabular datasets compared to unregularized gradient boosting.

D. Evaluation Metrics

Model performance is quantified using five complementary metrics. Accuracy measures the proportion of correctly classified samples:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision measures the proportion of positive predictions that are correct:

$$\text{Precision} = TP / (TP + FP)$$

Recall measures the proportion of actual positive cases that are correctly identified:

$$\text{Recall} = TP / (TP + FN)$$

F1-Score is the harmonic mean of Precision and Recall, providing a single metric that balances both:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

ROC-AUC summarizes classifier discrimination ability across all possible classification thresholds by computing the area under the Receiver Operating Characteristic curve, where a value of 1.0 indicates perfect discrimination and 0.5 indicates random performance.

IV. RESULTS AND DISCUSSION

A. Full Model Comparison

Table II presents the complete evaluation results for all three models across all four disease datasets on the held-out test set.

Disease	Model	Accuracy	Precision	Recall	F1	ROC-AUC	CV Mean Acc
Diabetes	Logistic Regression	0.7078	0.5714	0.6667	0.6154	0.8085	0.7589
Diabetes	Random Forest	0.7208	0.5733	0.7963	0.6667	0.8248	0.7638
Diabetes	XGBoost	0.7403	0.5921	0.8333	0.6923	0.8167	0.7589
Heart Disease	Logistic Regression	0.8500	0.8519	0.8214	0.8364	0.9498	0.8143
Heart Disease	Random Forest	0.8667	0.8846	0.8214	0.8519	0.9408	0.8099
Heart Disease	XGBoost	0.8333	0.8750	0.7500	0.8077	0.9330	0.8099
Liver Disease	Logistic Regression	0.7094	0.9153	0.6506	0.7606	0.8398	0.6288
Liver Disease	Random Forest	0.7265	0.8493	0.7470	0.7949	0.8047	0.6523
Liver Disease	XGBoost	0.7350	0.7766	0.8795	0.8249	0.7562	0.6931
Chronic Kidney Disease	Logistic Regression	0.9750	1.0000	0.9600	0.9796	0.9993	0.9938
Chronic Kidney Disease	Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	0.9906
Chronic Kidney Disease	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	0.9812

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

B. Best Model Per Disease

Table III. Best Performing Model Per Disease Dataset

Disease	Best Model	Accuracy	Precision	Recall	F1	ROC-AUC	CV Mean Acc
Diabetes	Random Forest	0.7208	0.5733	0.7963	0.6667	0.8248	0.7638
Heart Disease	Logistic Regression	0.8500	0.8519	0.8214	0.8364	0.9498	0.8143
Liver Disease	Logistic Regression	0.7094	0.9153	0.6506	0.7606	0.8398	0.6288
Chronic Kidney Disease	Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	0.9906

C. Discussion

For diabetes prediction, XGBoost achieves the highest test accuracy (0.7403) and recall (0.8333), indicating stronger sensitivity to positive cases. Random Forest achieves the highest ROC-AUC (0.8248) and CV mean accuracy (0.7638), suggesting better generalization stability across folds. In clinical screening contexts where false negatives carry a higher cost than false positives, the superior recall of XGBoost may make it the preferred choice despite its slightly lower ROC-AUC compared to Random Forest.

For heart disease prediction, Logistic Regression achieves the highest ROC-AUC (0.9498) and CV mean accuracy (0.8143), which is a notable finding. This result suggests that the heart disease feature set contains strong linear discriminative signals that Logistic Regression captures efficiently. The relatively small dataset size of 297 samples may also contribute to marginal overfitting in the more complex ensemble models, explaining their slightly lower cross-validation performance compared to the linear baseline.

For liver disease prediction, XGBoost achieves the highest recall (0.8795) and F1-score (0.8249), making it the most practically useful model for this heavily imbalanced dataset. The moderate cross-validation mean accuracy of 0.6931 reflects the inherent difficulty of liver disease classification from the available features and suggests that additional clinical features or external validation data may be needed to build a more reliable predictor.

For chronic kidney disease, both Random Forest and XGBoost achieve perfect scores across all evaluation metrics on the test set. Random Forest shows a slightly higher CV mean accuracy (0.9906) compared to XGBoost (0.9812). The high discriminative power of features such as hemoglobin and serum creatinine likely contributes to this result. Researchers should conduct careful feature analysis to verify that near-perfect performance does not reflect dataset-specific properties such as near-zero-variance features or near-perfect linear separability

V. EXPLAINABILITY ANALYSIS

SHAP summary plots, bar plots, and dependence plots are generated for XGBoost models on all four disease datasets. Feature importance plots are additionally computed for Logistic Regression using coefficient magnitude and for Random Forest using mean decrease in impurity.

For the diabetes dataset, glucose concentration emerges as the dominant feature driving positive predictions, followed by BMI and diabetes pedigree function. This is consistent with established clinical understanding of type 2 diabetes risk factors, where elevated blood glucose is the primary diagnostic indicator and obesity measured by BMI is a major modifiable risk factor.

For heart disease prediction, chest pain type and maximum heart rate achieved during exercise testing show the highest SHAP importance values. The negative association between maximum heart rate and disease risk, where lower maximum heart rate corresponds to higher disease probability, aligns with the known clinical relationship between reduced exercise tolerance and cardiac dysfunction.

DOI: 10.48175/IJARSCT-35427

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

For liver disease prediction, total bilirubin and direct bilirubin levels contribute most strongly to positive predictions. Elevated bilirubin is a well-established clinical marker of liver function impairment, providing face validity to the model's learned feature importance

VI. LIMITATIONS

Several limitations of this benchmark study must be acknowledged. The datasets used are publicly available benchmark datasets that may not represent the demographic and clinical characteristics of specific local patient populations. Results should not be generalized to populations with different epidemiological profiles without external validation.

The perfect classification results observed on the chronic kidney disease dataset warrant cautious interpretation. Near-perfect performance on small datasets can reflect dataset-specific properties such as highly discriminative individual features rather than genuine model generalization capability. External validation on an independent CKD cohort is necessary before drawing strong conclusions about model utility.

The current framework does not include probability calibration analysis. Raw predicted probabilities from XGBoost and Random Forest may not accurately reflect true event rates, which is an important consideration for clinical risk communication. No confidence intervals are reported for the evaluation metrics, and bootstrap-based uncertainty quantification should be added in future work.

SMOTE-generated synthetic samples are based on interpolation within the training distribution and may not accurately represent the true underlying distribution of minority-class patients in clinical practice. The framework uses public benchmark datasets that have known limitations including zero values for physiologically implausible features such as zero glucose or zero BMI in the diabetes dataset, which should be addressed through domain-specific preprocessing in future iterations.

VII. FUTURE WORK

Several directions are identified for extending this benchmark study. External validation on independent clinical datasets collected from different institutions and patient populations is the most important next step for assessing true generalization capability.

Probability calibration using Platt scaling or isotonic regression should be incorporated to ensure that predicted risk scores accurately reflect empirical event rates. Bootstrap-based confidence intervals should be computed for all reported metrics to quantify uncertainty and support statistical comparison between models.

Fairness analysis across demographic subgroups including age groups, sex, and where available ethnicity is an important requirement for responsible deployment of clinical prediction models. Performance disparities across subgroups can have significant consequences in healthcare settings and should be explicitly evaluated and reported.

VIII. CONCLUSION

This paper presented a benchmark study evaluating Logistic Regression, Random Forest, and XGBoost for predicting diabetes, heart disease, liver disease, and chronic kidney disease under a unified experimental framework. The framework applies leakage-safe preprocessing, SMOTE-based class balancing within cross-validation pipelines, GridSearchCV hyperparameter tuning, standardized metric reporting, and SHAP-based post-hoc explainability across all four disease datasets.

Experimental results demonstrate that no single model achieves superior performance across all datasets, reinforcing the importance of empirical model selection for each specific prediction task. Gradient boosting and ensemble methods show strong recall performance on imbalanced datasets such as diabetes and liver disease, while Logistic Regression remains competitive for heart disease prediction where linear feature relationships are dominant. Both Random Forest and XGBoost achieve perfect classification on the chronic kidney disease dataset, though this result requires careful validation before clinical interpretation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, May 2026
Impact Factor: 8.2

SHAP analysis confirms that the most influential features identified by the models align closely with established clinical knowledge across all four disease domains, supporting the interpretability and trustworthiness of the predictions. The complete framework with saved model artifacts, evaluation metrics, and explainability outputs provides a reproducible foundation for future research in clinical machine learning.

ACKNOWLEDGMENT

I would like to sincerely thank **Hemant Kumar, Associate Professor, Department of Computer Science and Engineering, Raffles University**, for his valuable guidance, continuous support, and helpful suggestions throughout this project.

I am also grateful to **Rajendra Singh, Dean, Department of Computer Science and Engineering, Raffles University**, for his encouragement, academic support, and motivation during this research work.

REFERENCES

[1] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

[2] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease

[3] Ramana, B. V., & Venkateswarlu, N. B. (2012). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. DOI: 10.24432/C5D02C.

[4] Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Dataset. UCI Machine Learning Repository. DOI: 10.24432/C5G020.

[5] 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. DOI: 10.1145/2939672.2939785

[6] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. DOI: 10.1613/jair.953

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[9] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. DOI: 10.1038/s42256-019-0138-9

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[11] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1–5.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. DOI: 10.1214/aos/1013203451

[13] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[14] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. DOI: 10.1007/978-0-387-84858-7

[15] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861 874. DOI: 10.1016/j.patrec.2005.10.010