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 5, May 2026
Impact Factor: 8.2

# A Reproducible Multi-Disease Prediction Framework Using Ensemble and Gradient Boosting Models with SHAP Explainability

**Tanmay Yadav[1], Hemant Kumar[2], Rajendra Singh[3]**
[1] Department of Computer Science and Engineering
[2] Assistant Professor, Department of Computer Science and Engineering
[3] Dean, Department of Computer Science and Engineering
Raffles University, Neemrana, Rajasthan, India
tanmayyadav300@gmail.com, hemantkumar@rafflesuniversity.edu.in
rajendra.singh@rafflesuniversity.edu.in

**Abstract:** *This paper presents a reproducible multi-disease prediction framework targeting diabetes, heart disease, liver disease, and chronic kidney disease. The framework systematically compares Logistic Regression, Random Forest, and XGBoost classifiers using leakage-safe preprocessing pipelines, SMOTE-based class balancing, 5-fold stratified cross-validation, GridSearchCV hyperparameter tuning, and SHAP-based explainability analysis. Experimental results demonstrate that XGBoost and Random Forest consistently outperform Logistic Regression across most disease datasets, with Random Forest and XGBoost achieving perfect classification metrics on the chronic kidney disease dataset. The system is designed as an educational research implementation and does not constitute a clinical diagnostic tool*

**Keywords:** *multi-disease*

## I. INTRODUCTION

Early detection of chronic diseases such as diabetes, heart disease, liver disease, and chronic kidney disease remains a critical challenge in healthcare systems worldwide. Machine learning methods have demonstrated substantial capability in supporting early screening by learning complex patterns from structured clinical datasets. Unlike deep learning approaches that require large unstructured datasets, classical machine learning models trained on tabular clinical data offer interpretability, reproducibility, and practical deployment potential in resource-constrained environments.

Existing studies commonly evaluate individual models or compare a limited subset of algorithms on a single disease dataset. This work addresses that gap by presenting a unified multi-disease prediction framework that applies identical preprocessing and evaluation protocols across four clinically relevant disease domains. The system is designed for reproducibility, with fixed random seeds, pipeline-safe SMOTE application, and standardized metric reporting.

The primary contributions of this paper are as follows. First, a leakage-safe preprocessing pipeline is implemented that applies SMOTE strictly within the training fold during cross-validation. Second, a systematic three-model comparison is conducted across four disease datasets under identical experimental conditions. Third, SHAP-based explainability is integrated for XGBoost models to provide feature-level interpretability. Fourth, all models, scalers, and evaluation artifacts are saved for reproducibility and downstream deployment.

## II. LITERATURE REVIEW

Prior work in machine learning-based medical prediction has extensively evaluated linear models, tree ensembles, and gradient boosting methods on structured patient data. Logistic Regression remains a strong baseline due to its

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 5, May 2026
Impact Factor: 8.2

probabilistic output and interpretability. Random Forest improves non-linear robustness through bootstrap aggregation of decision trees, reducing variance while maintaining reasonable bias. XGBoost has demonstrated superior performance on tabular medical datasets through regularized gradient boosting, which sequentially reduces prediction error while penalizing model complexity.

Studies on diabetes prediction using the Pima Indians Diabetes Database have reported accuracy ranges between 0.70 and 0.80 for classical models. Heart disease prediction using the Cleveland dataset has yielded AUC values above 0.90 for ensemble methods. Liver disease prediction presents additional challenges due to class imbalance, where precision and recall trade-offs become significant. Chronic kidney disease datasets, characterized by a high number of features and relatively clean structure, have consistently produced near-perfect classification results with tree-based models.

Class imbalance handling through SMOTE has been widely shown to improve recall for minority classes in medical datasets. However, a common methodological error in prior work is applying SMOTE before cross-validation splitting, which causes data leakage from validation folds into synthetic training samples. The present framework addresses this by encapsulating SMOTE within the training pipeline using imbalanced-learn's Pipeline construct.

SHAP (SHapley Additive exPlanations) has emerged as the standard method for post-hoc model explainability in tree-based models. SHAP values decompose individual predictions into feature-level contributions, enabling clinicians and researchers to understand which input variables most influence a given prediction.

## III. METHODOLOGY

### A. Dataset Summary

Four publicly available clinical datasets are used in this study. The Pima Indians Diabetes dataset contains 768 samples with 8 features and a binary outcome indicating diabetes presence. The Heart Disease dataset from the UCI Machine Learning Repository contains 297 samples with 13 features. The Indian Liver Patient dataset contains 583 samples with 10 features. The Chronic Kidney Disease dataset contains 400 samples with 24 features including both numerical and categorical attributes.

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

### B. Preprocessing Pipeline

All datasets undergo a standardized preprocessing workflow. Missing values are imputed using median imputation to avoid mean-shift bias from outliers. Categorical features are encoded using label encoding before model training. Feature standardization is applied using StandardScaler, which transforms each feature to zero mean and unit variance according to:

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

where $\mu$ is the training mean and $\sigma$ is the training standard deviation. Critically, the scaler is fit exclusively on training data and applied to validation and test data to prevent leakage.

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 5, May 2026 Impact Factor: 8.2

Feature engineering includes interaction terms and domain-relevant derived features where applicable. SMOTE oversampling is applied inside the training pipeline to generate synthetic minority-class samples, ensuring that no synthetic samples influence validation fold metrics during cross-validation.

**C. Model Descriptions**

Three model families are evaluated. Logistic Regression estimates the probability of the positive class using the sigmoid function:

$P(y=1|x) = 1 / (1 + e^{(-z)})$, where $z = w^T x + b$

The decision boundary is the hyperplane where the predicted probability equals the selected classification threshold.

Random Forest is an ensemble of decision trees trained on bootstrap samples. Node splitting uses Gini impurity:

$Gini = 1 - \Sigma(p_i^2)$

where $p_i$ is the proportion of class i in a node. Final classification is determined by majority vote across all trees.

XGBoost performs gradient boosting by sequentially adding trees to correct residual errors of previous iterations. Its regularized objective function is:

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

where the regularization term $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ penalizes tree complexity through leaf count T and weight magnitude, improving generalization on tabular data.

**D. Experimental Setup**

All experiments use random_state=42 and np.random.seed(42) for full reproducibility. Model selection is performed using 5-fold stratified cross-validation, which preserves class proportions across folds. Hyperparameter tuning is conducted using GridSearchCV with cross-validated scoring on F1 to account for class imbalance. SMOTE is applied exclusively within the training fold at each cross-validation iteration using imbalanced-learn's Pipeline construct.

**E. Evaluation Metrics**

Model performance is evaluated using five metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

$Precision = TP / (TP + FP)$

$Recall = TP / (TP + FN)$

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

ROC-AUC summarizes discrimination capability across all classification thresholds using the area under the Receiver Operating Characteristic curve. Cross-validation mean accuracy is additionally reported to assess generalization stability.

**F. Explainability**

SHAP summary plots, bar plots, and dependence plots are generated for XGBoost models on each disease dataset. Feature importance plots are saved for all three model families where supported by the respective implementation. SHAP analysis allows identification of the most influential clinical features driving individual predictions.

## IV. RESULTS AND DISCUSSION

**A. Full Model Comparison**

The table below presents evaluation metrics for all three models across four disease datasets on the held-out test set.

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 5, May 2026
Impact Factor: 8.2

| 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 |

**C. Discussion**

For diabetes prediction, XGBoost achieves the highest test accuracy (0.7403) and recall (0.8333), indicating stronger sensitivity to positive cases, while Random Forest achieves the highest ROC-AUC (0.8248) and CV mean accuracy (0.7638), suggesting better generalization. In clinical screening contexts where false negatives carry higher cost than false positives, the higher recall of XGBoost may be preferred.

For heart disease prediction, Logistic Regression achieves the highest ROC-AUC (0.9498) and CV mean accuracy (0.8143), which is a notable result. This suggests that the heart disease feature set contains strong linear signals that Logistic Regression captures effectively, while ensemble methods may introduce marginal overfitting on the relatively small dataset of 297 samples.

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 imbalanced dataset. However, cross-validation mean accuracy (0.6931) remains moderate, reflecting the inherent difficulty of liver disease classification from the available features.

For chronic kidney disease, both Random Forest and XGBoost achieve perfect scores across all metrics, with Random Forest showing slightly higher CV mean accuracy (0.9906 vs 0.9812). The clean structure and high feature count of the CKD dataset likely contribute to this result. Researchers should verify that this does not reflect dataset-level issues such as near-perfect feature separation before reporting in clinical contexts.

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 5, May 2026
Impact Factor: 8.2

## V. EXPLAINABILITY ANALYSIS

SHAP summary, bar, and dependence plots are generated for XGBoost models on all four disease datasets. For the diabetes dataset, glucose concentration and BMI emerge as the dominant features driving positive predictions, consistent with clinical understanding of diabetes risk factors. For heart disease, features related to chest pain type and maximum heart rate show the highest SHAP importance. For chronic kidney disease, hemoglobin and serum creatinine are identified as the most influential features.

Feature importance plots are additionally saved for Logistic Regression (using coefficient magnitude) and Random Forest (using mean decrease in impurity). These artifacts support post-hoc analysis and can be reviewed alongside SHAP outputs for cross-method validation of feature relevance.

## VI. LIMITATIONS

Several limitations of this study must be acknowledged. The datasets used are publicly available benchmark datasets sourced primarily from the UCI Machine Learning Repository. These datasets may not represent the demographic and clinical characteristics of specific local patient populations, limiting direct generalizability. The chronic kidney disease perfect classification results warrant cautious interpretation, as near-perfect separation in small datasets can reflect dataset-specific properties rather than genuine model capability.

The implementation is not clinically validated and must not be used as a diagnostic device. No confidence interval analysis or calibration assessment is included in this study. Class imbalance handling through SMOTE generates synthetic samples based on the training distribution, which may not accurately reflect real minority-class cases in unseen populations.

## VII. FUTURE WORK

Future extensions of this framework should include external validation on independent clinical datasets to assess true generalization capability. Probability calibration analysis using Platt scaling or isotonic regression should be incorporated to ensure that model confidence scores are reliable for clinical interpretation. Bootstrap-based confidence intervals should be computed for all reported metrics to quantify uncertainty in performance estimates.

Fairness analysis across demographic subgroups is an important direction, as model performance disparities across age, sex, and ethnicity can have significant implications in healthcare applications. Prospective data collection and collaboration with clinical teams for feature interpretation review would substantially strengthen the translational relevance of this framework. Integration of additional model families such as LightGBM, CatBoost, and neural network baselines would provide a more comprehensive comparative study.

## VIII. CONCLUSION

This paper presented a reproducible IEEE-style experimental framework for multi-disease prediction covering diabetes, heart disease, liver disease, and chronic kidney disease. The framework applies leakage-safe preprocessing, SMOTE-based class balancing within cross-validation pipelines, GridSearchCV hyperparameter tuning, and SHAP-based explainability across three model families. Experimental results demonstrate that no single model dominates across all disease datasets, reinforcing the importance of dataset-specific model selection. The framework provides full evaluation metrics, cross-validation results, saved model artifacts, and explainability assets to support reproducible research in clinical machine learning.

**Acknowledgment**

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

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 5, May 2026
Impact Factor: 8.2

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