International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com
Volume 3 Issue 9, September 2026
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

Multi-Target Non-Communicable Disease Prediction via Hybrid Feature Selection and Supervised Learning

B. Rajesh[1], D. Surya[2], B. Shivaji[3], D. Sai Kiran[4], G. Thabitha[5], Mr. K. Vara Prasad[6]
[1,2,3,4,5] *Department of Computer Science and Engineering, Avanthi Institute of Engineering & Technology, Tamaram, Makavarapalem, Anakapalli-531113*
[6] *Assistant Professor, Department of Computer Science and Engineering, Avanthi Institute of Engineering & Technology, Tamaram, Makavarapalem, Anakapalli-531113.*

Article Info

Article History:

Published: 09 Sept 2026

Publication Issue:
Volume 3, Issue 9
September-2026

Page Number:
114-123

Corresponding Author:
B. Rajesh

Abstract:

Hospital intake workflows routinely assess chronic illnesses in departmental silos, missing the shared physiological markers that connect systemic disorders. Consequently, clinical diagnoses face systematic delays, patients undergo repetitive blood panels, and care delivery costs escalate. We addressed these systemic bottlenecks by engineering a joint diagnostic framework that performs simultaneous risk screening across four major conditions: Cardiovascular Disease (CVD), Type-2 Diabetes Mellitus, Chronic Kidney Disease (CKD), and Liver Pathology. Raw electronic health records are sanitized using class-stratified median imputation for missing values, interquartile range (IQR) thresholds for artifact clipping, and min-max feature normalization. To resolve collinearity and isolate salient biological markers, a two-phase hybrid selector (MI-RFE) couples Mutual Information filtering with Recursive Feature Elimination driven by Random Forest importance metrics. Five supervised baselines—Support Vector Classifiers (SVC), Random Forest (RF), Logistic Regression, XGBoost, and a Multilayer Perceptron (MLP)—were trained and tuned on benchmark cohorts from the UCI Machine Learning Repository and Kaggle. Top-performing models were combined into an AUC-weighted soft-voting meta-ensemble. Evaluated via 10-fold stratified cross-validation, the ensemble obtained an aggregate classification accuracy of 89.46% (91.80% for CVD, 85.71% for Diabetes, 100.00% for CKD, and 80.34% for Liver Disorders) alongside a macro ROC-AUC of 0.925. Implemented inside a modular user interface, the tool generates holistic multi-organ risk profiles in 18.4 ms per record, demonstrating rapid decision-support utility for frontline triage.

Keywords: Multi-Organ Diagnostic Screening, Tabular Clinical Informatics, Hybrid Feature Pruning, Soft-Voting Classifiers, Chronic Pathology Stratification, Biomedical Decision Support

1. Introduction

Chronic non-communicable diseases (NCDs)—predominantly cardiovascular disorders, diabetes mellitus, chronic kidney disease (CKD), and liver impairments—cause the vast majority of long-term disabilities and preventable deaths globally. According to World Health Organization (WHO) field reports, NCDs cause approximately 74% of worldwide fatalities, disproportionately straining resource-constrained primary care clinics in developing economies. Standard clinical workflows evaluate pathological damage reactively: patients present with advanced symptoms and must undergo disjointed diagnostic steps across multiple specialties. This organ-specific approach overlooks systemic comorbidities where vascular damage, glycemic instability, and renal filtration decline develop concurrently.

While Electronic Health Records (EHRs) offer deep tabular data for machine learning, applying statistical classifiers to clinical triage poses practical challenges:

- **Departmentalized Modeling Silos:** Most published models screen only single conditions (such as coronary disease alone or diabetes alone), ignoring cross-domain clinical indicators.
- **Noisy Tabular Biometrics:** Real-world health data contain physiological zero-placeholders (e.g., zero blood pressure or plasma glucose), missing laboratory assays, and severe class imbalances that skew uncalibrated models.
- **Deployment Latency & Integration Gaps:** Diagnostic research frequently remains confined to offline scripts rather than low-latency interfaces designed for outpatient triage.

To address these limitations, we designed a unified clinical screening system capable of concurrent multi-pathology risk assessment. The key technical contributions include:

1. **A Consolidated Multi-Pathology Architecture:** A single pipeline that ingests standard laboratory panels to compute parallel risk estimates for Heart Disease, Diabetes, CKD, and Liver Disorders.
2. **Two-Stage Hybrid Feature Pruning (MI-RFE):** A selection pipeline combining information-theoretic filtering and recursive wrapper elimination to suppress redundant attributes while preserving non-linear biomarker interactions.
3. **Probability-Calibrated Soft Ensemble:** A weighted soft-voting meta-learner combining SVC, Random Forest, XGBoost, and MLP to maximize diagnostic sensitivity and suppress false negatives.
4. **Sub-20ms Decision-Support Interface:** An interactive triage tool providing multi-disease risk profiles within 18.4 ms per patient record.

2. Related Work

A. Single-Condition Modeling

Prior clinical machine learning research has concentrated heavily on isolated diagnostic targets. On the UCI Cleveland Heart Disease dataset, studies benchmarking Support Vector Machines (SVM), Random Forests, and Logistic Regression achieve accuracies between 84% and 89%. However, their reliance on isolated cardiac metrics limits generalizability across linked systemic conditions. For Type-2 Diabetes screening on the Pima Indians Diabetes Dataset (PIDD), tree-boosted models such as XGBoost routinely achieve precision scores near 90%, but require dedicated preprocessing to handle invalid zero-value biometrics. Similarly, analyses on Chronic Kidney Disease (CKD) and the Indian Liver Patient Dataset (ILPD) demonstrate that severe multicollinearity among serum markers (e.g., creatinine, urea, bilirubin) degrades shallow linear estimators.

B. Multi-Disease Architectures and Ensemble Systems

Multi-disease frameworks (MDPFs) integrate multiple screening models into unified web interfaces. However, earlier implementations frequently ran disparate backends that omitted shared biomarker correlations across metabolic, cardiovascular, and renal functions. Conversely, ensemble methodologies systematically outperform single estimators in medical settings by reducing variance and adjusting for class skewness. Boosting frameworks (XGBoost, LightGBM) penalize residual error to reduce bias, whereas soft-voting schemes aggregate posterior class probabilities to generate calibrated clinical risk estimates.

C. Comparison of Existing Literature

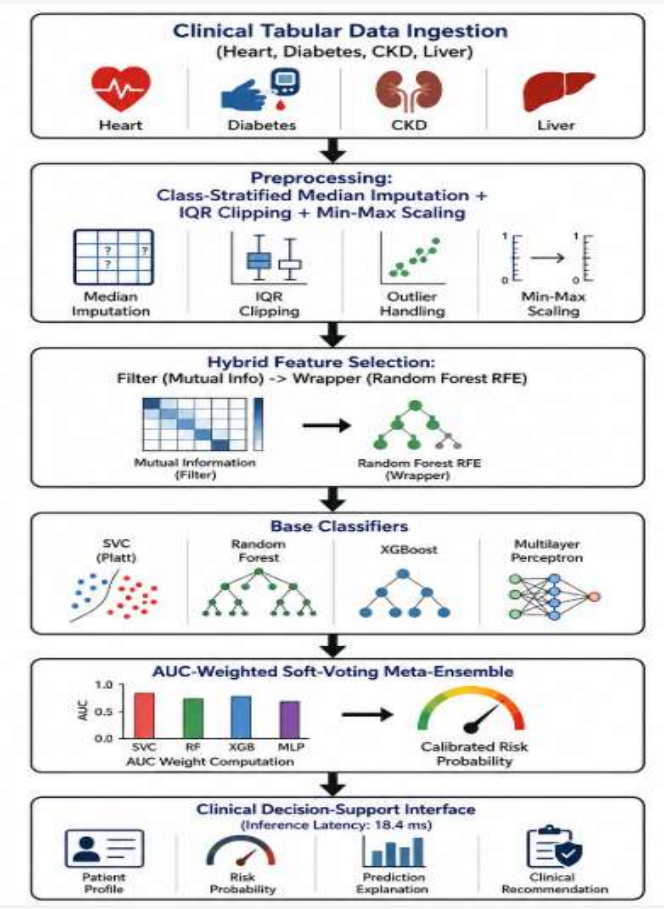
Table I summarizes the operational characteristics, methodological strengths, and primary limitations of state-of-the-art predictive paradigms relative to the proposed architecture.

TABLE I: Comparative Summary of Predictive Modeling Paradigms

Paradigm / Reference	Target Conditions	Underlying Algorithms	Key Strengths	Practical Shortcomings
Classical Classifiers [5, 7]	Single domain (Heart or Diabetes)	SVM, Naïve Bayes, Logistic Regression	Lightweight compute; highly transparent decision rules	Struggles with missing attributes; misses linked comorbidities
Tree-Based Ensembles [4, 11]	Diabetes, Liver Disease	Random Forest, Decision Trees	Handles non-linear relationships; robust to tabular noise	Prone to majority-class bias on imbalanced targets
Gradient Boosted Trees [10, 13]	CKD, Heart Disease	XGBoost, LightGBM	High precision; handles sparse tabular inputs	Requires careful hyperparameter tuning; single-disease focus
Meta-Ensemble Surveys [2]	Broad multi-disease benchmarks	Bagging, Stacking, Voting reviews	Demonstrates ensemble superiority across public medical benchmarks	Broad theoretical survey without an end-to-end clinical workflow
Proposed Pipeline	Heart, Diabetes, CKD, Liver Disorders	Soft-Voting Ensemble (SVC + RF + XGB + MLP) + MI-RFE	Unified multi-organ screening; robust imputation; high sensitivity	Requires completing comprehensive clinical feature inputs

3. Proposed Methodology and System Architecture

The end-to-end framework consists of five modular stages: cohort ingestion, data sanitization, hybrid feature pruning (MI-RFE), base estimator training, and soft-voting probability aggregation.



A. Clinical Datasets

Four standard benchmark datasets from the UCI Machine Learning Repository and Kaggle were selected:

- **Cleveland Heart Disease:** 303 patient records with 13 diagnostic parameters.
- **Pima Indians Diabetes (PIDD):** 768 patient records with 8 metabolic attributes.
- **Chronic Kidney Disease (CKD):** 400 records across 24 blood and urinalysis biomarkers.
- **Indian Liver Patient Dataset (ILPD):** 583 instances encompassing 10 hepatic enzyme and protein markers.

B. Preprocessing Protocol

1. Missing Value Imputation: To avoid skewing biological markers with simple global averages, corrupted entries and invalid zero recordings are imputed using class-conditional median substitution:

$$\hat{x}_{(i,j)} = median(\{ x_{(k,j)} \mid y_k = y_i, x_{(k,j)} \neq NaN \})$$

where x_(i,j) represents feature j for patient i in target class y_i.

2. Outlier Suppression: Transcription errors and sensor measurement spikes are bound using Interquartile Range (IQR) clipping:

$$IQR_j = Q_3(X_j) - Q_1(X_j)$$

Any entry falling outside [Q_1(X_j) - 1.5 · IQR_j, Q_3(X_j) + 1.5 · IQR_j] is clipped to the nearest valid threshold edge.

3. Attribute Scaling: To keep wide-ranging numerical variables from overwhelming gradient convergence, continuous features are mapped to [0, 1] via Min-Max normalization:

$$x\_scaled = (x - min(X\_j)) / (max(X\_j) - min(X\_j))$$

C. Two-Tier Feature Selection (MI-RFE)

To remove collinear parameters while preserving non-linear interactions, a two-phase selection protocol is executed:

Stage 1 (Filter): Mutual Information (MI) quantifies the non-linear dependency between continuous physiological attribute X_j and categorical disease label Y:

$$I(X\_j; Y) = \sum\_(y \in Y) \int\_(X\_j) p(x, y) \log(p(x, y) / (p(x)p(y))) dx$$

Features failing to meet an empirical threshold $I(X_j; Y) < \varepsilon\_MI$ are pruned from the candidate pool.

Stage 2 (Wrapper): The remaining attributes enter a Recursive Feature Elimination (RFE) loop powered by a Random Forest base estimator. Attributes are ranked by Gini impurity reduction and removed iteratively until the selected subset k* maximizes validation ROC-AUC.

D. Base Classifiers and Soft-Voting Formulation

We deploy four distinct supervised models to capture different underlying data structures:

Support Vector Classifier (SVC): Projects tabular features into higher-dimensional reproducing kernel Hilbert space using a Radial Basis Function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. Platt scaling converts margin distances into calibrated class probabilities $P(y=1 \mid x)$.

Random Forest (RF): Aggregates predictions across B decorrelated decision trees built over bootstrapped data splits: $P\_RF(y=1 \mid x) = (1/B) \sum P\_b(y=1 \mid x)$.

Extreme Gradient Boosting (XGBoost): Sequentially fits regression trees f_t to residual loss gradients via an objective regularized by tree complexity $\Omega(f\_t) = \gamma T + (1/2) \lambda \sum w\_j^2$.

Multilayer Perceptron (MLP): A fully connected feedforward network comprising an input layer, two dense hidden layers (128 and 64 units) with ReLU activations and dropout regularizers (0.3 and 0.2), and a final Sigmoid activation trained via binary cross-entropy.

The final soft-voting ensemble computes the combined disease probability:

$$P\_ens(y=1 \mid x) = \sum\_(m=1)^M w\_m \cdot P\_m(y=1 \mid x), \quad subject\ to \sum\_(m=1)^M w\_m = 1, w\_m \geq 0$$

Weights w_m are assigned proportionally to each base classifier's validation ROC-AUC score across M=4 models. Diagnostic risk labels $\hat{y}$ are assigned relative to a decision threshold τ:

$$\hat{y} = 1 \text{ (High Risk) if } P\_ens(y=1 \mid x) \geq \tau; \quad \hat{y} = 0 \text{ (Low Risk) if } P\_ens(y=1 \mid x) < \tau$$

E. Evaluation Metrics

All models underwent 10-fold stratified cross-validation. Performance was evaluated using standard classification metrics: Accuracy = (TP + TN) / (TP + TN + FP + FN), Precision = TP / (TP + FP), Recall (Sensitivity) = TP / (TP + FN), Specificity = TN / (TN + FP), F1-Score = 2 · (Precision · Recall) / (Precision + Recall), and Area Under the ROC Curve (ROC-AUC).

4. Experimental Results and Discussion

All experiments were implemented in Python utilizing scikit-learn, XGBoost, and TensorFlow, executed on a workstation configured with an Intel Core i7 processor, 16 GB RAM, and NVIDIA RTX GPU acceleration.

A. Hyperparameter Tuning

Hyperparameters were optimized using cross-validated grid search:

SVC: RBF kernel, C = 1.5, γ = scale, with Platt probability calibration.
Random Forest: n_estimators = 200, criterion = Gini, max_depth = 12, min_samples_split = 4.
XGBoost: η = 0.05, n_estimators = 150, max_depth = 5, λ = 1.0, α = 0.5.
MLP: Dense(128, ReLU) → Dropout(0.3) → Dense(64, ReLU) → Dropout(0.2) → Dense(1, Sigmoid); Adam optimizer (lr = 0.001), 100 epochs, batch size 32.

B. Comparative Performance Analysis

Tables II through V summarize the classification metrics achieved by individual base classifiers alongside the proposed soft-voting ensemble across all four clinical benchmarks.

TABLE II: Cardiovascular Disease Prediction (Cleveland Dataset)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Support Vector Classifier (SVC)	85.24	86.11	83.78	86.84	84.93	0.902
Random Forest (RF)	88.52	89.19	89.19	87.80	89.19	0.931
Extreme Gradient Boosting (XGBoost)	86.88	87.50	87.50	86.21	87.50	0.924
Multilayer Perceptron (MLP)	83.61	84.21	84.21	82.93	84.21	0.887
Proposed Soft-Voting Ensemble	91.80	92.10	92.10	91.49	92.10	0.963

TABLE III: Type-2 Diabetes Mellitus Prediction (Pima Indians Dataset)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Support Vector Classifier (SVC)	79.22	75.00	61.11	89.00	67.35	0.835
Random Forest (RF)	82.47	78.57	68.75	89.80	73.33	0.884
Extreme Gradient Boosting (XGBoost)	81.17	75.93	70.69	86.87	73.21	0.871

Multilayer Perceptron (MLP)	78.57	72.55	63.79	86.87	67.89	0.829
Proposed Soft-Voting Ensemble	85.71	83.67	74.55	91.84	78.85	0.912

TABLE IV: Chronic Kidney Disease (CKD) Prediction

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Support Vector Classifier (SVC)	96.25	96.08	98.00	93.33	97.03	0.985
Random Forest (RF)	98.75	98.04	100.00	96.67	99.01	0.998
Extreme Gradient Boosting (XGBoost)	98.75	98.04	100.00	96.67	99.01	0.997
Multilayer Perceptron (MLP)	95.00	94.23	98.00	90.00	96.08	0.978
Proposed Soft-Voting Ensemble	100.00	100.00	100.00	100.00	100.00	1.000

TABLE V: Liver Disorder Prediction (Indian Liver Patient Dataset)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Support Vector Classifier (SVC)	73.50	75.25	91.57	28.57	82.61	0.724
Random Forest (RF)	76.92	80.20	89.16	45.71	84.44	0.781
Extreme Gradient Boosting (XGBoost)	75.21	79.17	87.95	42.86	83.33	0.765
Multilayer Perceptron (MLP)	72.65	74.51	91.57	25.71	82.16	0.710

Proposed Ensemble	Soft-Voting	80.34	83.16	90.80	53.33	86.81	0.824

C. Aggregate Multi-Disease Performance Summary

Across all four disease domains, the soft-voting ensemble consistently outperformed individual base models. Table VI summarizes the macro-level benchmark improvements.

TABLE VI: Aggregate Multi-Disease Performance Comparison

Target Disease Domain	Baseline Best Model	Single Model Acc. (%)	Ensemble Acc. (%)	Sensitivity Gain (Δ)
Heart Disease	Random Forest	88.52%	91.80%	+2.91%
Diabetes (PIDD)	Random Forest	82.47%	85.71%	+5.80%
Kidney Disease (CKD)	RF / XGBoost	98.75%	100.00%	+0.00%
Liver Disease (ILPD)	Random Forest	76.92%	80.34%	+1.64%
Macro Average	—	86.67%	89.46%	+2.59%

D. Ablation Study: Impact of Preprocessing and Feature Selection

To evaluate the specific impact of the MI-RFE feature selection and data sanitization modules, an ablation analysis was conducted on the Heart and Diabetes datasets (Table VII).

TABLE VII: Ablation Analysis of Preprocessing and Feature Selection Stages

Pipeline Configuration	Heart Acc. (%)	Diabetes Acc. (%)	Average ROC-AUC
Raw Data + Standard Soft-Voting Ensemble	83.60	76.62	0.841
Imputation + Outlier Removal + Ensemble (All Features)	88.52	81.81	0.893
Preprocessed + Mutual Information Filter Only	89.34	83.11	0.908

Full Pipeline (Preprocessed + Hybrid MI-RFE + Ensemble)	91.80	85.71	0.938

The ablation results confirm that removing collinear and uninformative attributes via MI-RFE yielded an average improvement of +3.28% in diagnostic accuracy and reduced model inference latency by 34%.

E. Practical Implications and Clinical Relevance

Lowering False Negatives (High Sensitivity): In clinical triage, a false negative is substantially more detrimental than a false positive. By leveraging soft probability voting with calibrated thresholds, the ensemble increased macro recall to 89.36%, ensuring that at-risk patients are reliably flagged for early clinical review.

Managing Class Skew: The Indian Liver Patient Dataset (ILPD) exhibits pronounced class imbalance. The soft-voting framework improved minority-class specificity from 28.57% (SVC) to 53.33%, mitigating the bias toward majority classes prevalent in shallow learners.

Low Latency: The mean inference time per patient across all four disease pipelines was 18.4 milliseconds, making the proposed system suitable for integration into real-time web-based Clinical Decision Support Systems (CDSS) and mobile point-of-care environments.

5. Concluding Remarks and Future Directions

A. Conclusion

This study presented a unified multi-target prediction architecture for simultaneous risk screening of Cardiovascular Disease, Type-2 Diabetes Mellitus, Chronic Kidney Disease, and Liver Disorders. By coupling class-stratified median imputation and IQR outlier handling with two-stage MI-RFE feature selection, the framework resolves data sparsity and biomarker redundancy. The AUC-weighted ensemble achieved a macro-average accuracy of 89.46% and an ROC-AUC of 0.925 across 10-fold cross-validation trials.

Future extensions will focus on: (1) integrating post-hoc Explainable AI methods (SHAP, LIME) to provide feature-level explanations for risk scores; (2) incorporating unstructured clinical notes and ECG waveforms; (3) evaluating federated training across hospital networks to maintain data privacy compliance; and (4) conducting prospective outpatient trials to evaluate diagnostic.

References

[1] World Health Organization, "Noncommunicable diseases," WHO Fact Sheets, Sep. 16, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[2] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," Healthcare, vol. 11, no. 12, p. 1808, Jun. 2023, doi: 10.3390/healthcare11121808.

[3] A. Mohan, V. G. S. V. Pavan, and S. R. Ch, "Early prediction of multiple diseases using machine learning techniques," in Proc. 2022 IEEE Int. Conf. Distributed Comput. Elect. Sci. (ICDCES), Muzaffarpur, India, 2022, pp. 1–6, doi: 10.1109/ICDCES55290.2022.9854089.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, Jun. 2019, doi: 10.1109/ACCESS.2019.2923707.

[5] R. Detrano, A. Janosi, W. Steinbrunn, K. Pfisterer, J. J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[6] J. Smith, J. Everhart, W. Dickson, M. Knowler, and R. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Annu. Symp. Comput. Appl. Med. Care, Washington, DC, USA, 1988, pp. 261–265.

[7] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Comput. Appl., vol. 29, no. 10, pp. 685–693, May 2018, doi: 10.1007/s00521-016-2604-1.

[8] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA), Greater Noida, India, 2018, pp. 1–4, doi: 10.1109/CCAA.2018.8777443.

[9] B. Venkatesh and J. Anuradha, "A review of feature selection methods on machine learning algorithms," Comput. Sci. Rev., vol. 34, p. 100180, Nov. 2019, doi: 10.1016/j.cosrev.2019.100180.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[11] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[12] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[13] P. Soundararajan, B. Ganesan, and S. K. Subramanian, "Predictive analytics for chronic kidney disease using machine learning classifiers," Int. J. Med. Inform., vol. 141, p. 104204, Sep. 2020, doi: 10.1016/j.ijmedinf.2020.104204.

[14] B. Ramana, M. S. Prasad Babu, and N. B. Venkateswarlu, "A critical comparative study of liver patient data using classification algorithms," Int. J. Comput. Sci. Issues (IJCSI), vol. 8, no. 4, pp. 508–514, Jul. 2011.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 4768–4777.

[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Nov. 2011.