Disease Prediction Using Ensemble Learning and XAI: A Comparative Study on Heart, Diabetes, Kidney, Parkinson’s and Breast Cancer Datasets

Fouzia Zilani
Student ID: 2012176111
Session: 2019-20
Course: CSE4202
University of Rajshahi

Introduction

Early disease diagnosis saves lives and reduces healthcare costs.

Traditional diagnostic systems are **time-consuming**, **error-prone**, and **inconsistent**.

ML can detect hidden patterns in medical data.

The current study introduces an **ensemble-based predictive framework** supported by **Explainable Artificial Intelligence (XAI)** for five major diseases:

- Heart Disease
- Diabetes
- Chronic Kidney Disease (CKD)
- Parkinson's Disease
- Breast Cancer

The goal is to design a **transparent, generalizable, and accurate diagnostic system**.

Problem We Solved

Challenges in prior research:

- Models lacked consistency across datasets.
- Class imbalance reduced performance on disease-positive cases.
- Doctors could not interpret model’s decision.

Our Solution:

- Used **SMOTETomek** to balance data and remove noise.
- Built **Ensemble Model (Soft Voting)** combining Random Forest, XGBoost, LightGBM, CatBoost, SVM, and LR.
- Integrated **SHAP and LIME** to explain predictions.

Result: A *robust and interpretable framework* that performs well for **five major diseases.**

Research Objectives

1. Develop independent predictive models for five diseases.
2. Apply advanced data preprocessing:
    - Missing value handling
    - Feature scaling
    - Encoding categorical features
    - Data balancing (SMOTETomek)
3. Train and optimize base classifiers:
   Random Forest, XGBoost, LightGBM, CatBoost, SVM, and Logistic Regression.
4. Construct an **Ensemble Soft Voting Classifier** for improved robustness.
5. Evaluate models using **Accuracy, Precision, Recall, F1-Score**.
6. Apply **SHAP and LIME** for explainability of model predictions.

Workflow of the Proposed System

Dataset Collection
↓
Data Cleaning & Preprocessing
↓
SMOTETomek Balancing
↓
Feature Scaling
↓
Train Base Models
↓
Ensemble Voting
↓
Model Evaluation
↓
XAI (SHAP & LIME) Interpretation
↓
Model Evaluation & Deployment

Datasets Used

All datasets were obtained from **UCI Machine Learning Repository**:

Disease	Samples	Features	Type
Heart	303	13	Numeric + Categorical
Diabetes (PIMA)	768	8	Numeric
CKD	400	24	Mixed
Parkinson’s	197	23	Numeric
Breast Cancer	569	30	Numeric

Data Preprocessing

Steps implemented for all datasets:

1. **Missing Data Handling:**
   - Mean/median imputation for missing lab values (e.g., hemoglobin, albumin).
2. **Encoding:**
   - Label or One-Hot Encoding for categorical variables.
3. **Outlier Treatment:**
   - Threshold-based filtering to handle extreme medical readings.
4. **Feature Scaling:**
   - StandardScaler applied to normalize feature ranges.
5. **Train-Test Split:**
   - Stratified 80:20 split to maintain class balance.

Handling Class Imbalance – SMOTETomek

Medical datasets often have **fewer disease-positive cases** than healthy ones.

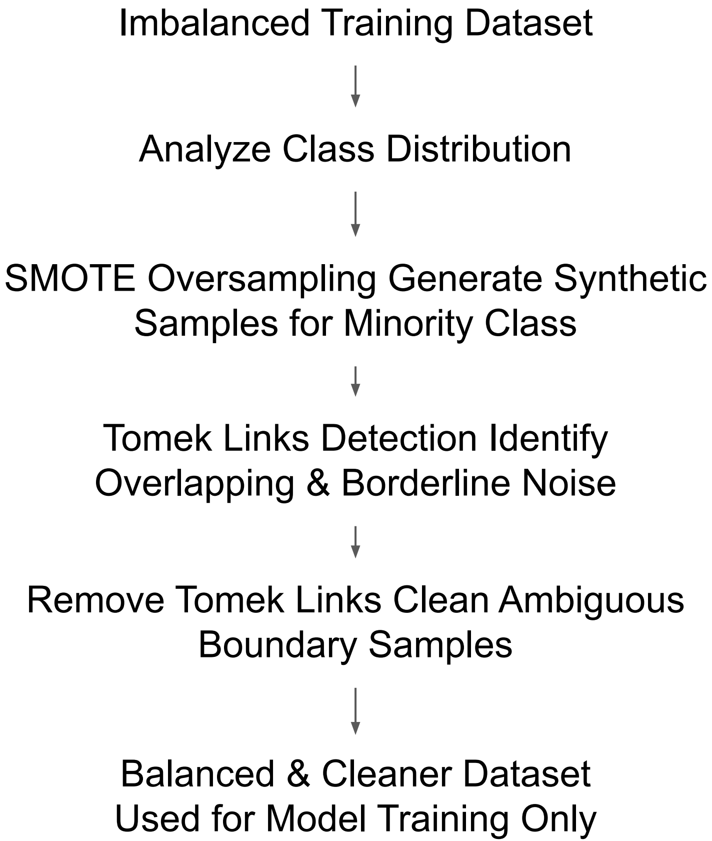
Used **SMOTETomek**, a hybrid method that:

- **SMOTE:** Creates synthetic minority samples.
- **Tomek Links:** Removes borderline and noisy samples.

Advantages:

- Prevents bias toward the majority class.
- Increases sensitivity (recall) for disease-positive samples.
- Produces cleaner decision boundaries for classifiers.

Pipeline for SMOTETomek-based class balancing:



Imbalanced Training Dataset

Analyze Class Distribution

SMOTE Oversampling Generate Synthetic Samples for Minority Class

Tomek Links Detection Identify Overlapping & Borderline Noise

Remove Tomek Links Clean Ambiguous Boundary Samples

Balanced & Cleaner Dataset Used for Model Training Only

Base Machine Learning Models

Six base classifiers were trained per disease dataset:

1. Logistic Regression
2. Random Forest
3. Support Vector Machine (SVM)
4. XGBoost
5. LightGBM
6. CatBoost

Each model tuned using grid search for optimal parameters.

CatBoost performed best with categorical data; XGBoost excelled in tabular numeric datasets.

Base models’ performance guided ensemble weighting.

Ensemble Learning Approach

Soft Voting Ensemble was implemented to integrate predictions from multiple models.
Formula:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \sum_{i=1}^{N} w_i \cdot P_i(c|x)$$

Where wi = weight of classifier i, Pi(c∣x) = predicted probability.

- Combines strengths of individual models → improved generalization.
- Reduces overfitting and variance.
- Achieves stable predictions across all five diseases.

Model Evaluation Metrics

To ensure clinically meaningful evaluation:

- **Accuracy:** Overall correctness.
- **Precision:** Reliability of positive predictions.
- **Recall (Sensitivity):** Ability to detect true disease cases.
- **F1-Score:** Balance between Precision & Recall.
- **Confusion Matrix:** Shows TP, FP, FN, TN — crucial in clinical evaluation.

In healthcare, **False Negatives (FN)** are more critical than False Positives.

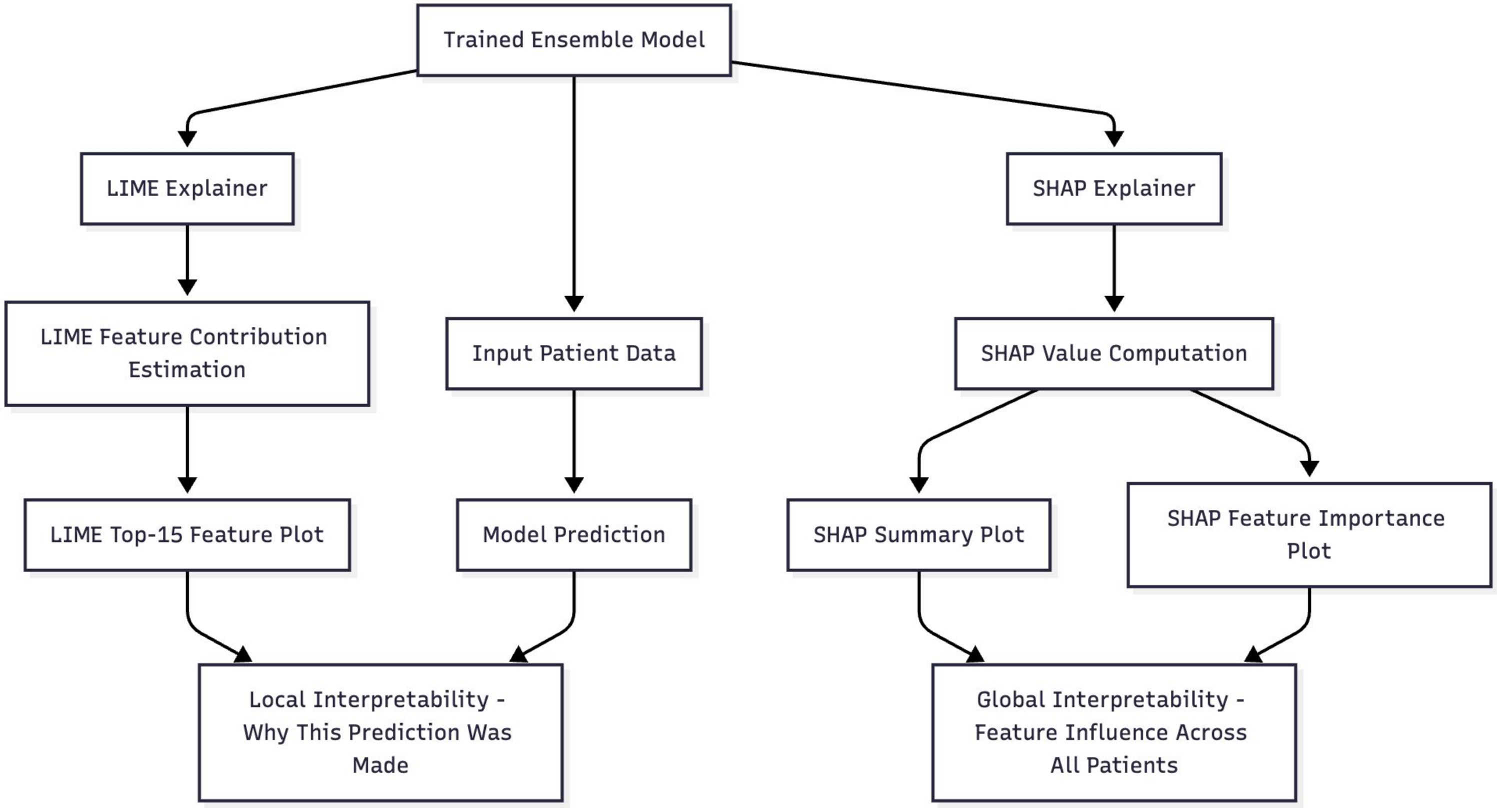
Explainable Artificial Intelligence (XAI)

To interpret model predictions, two XAI tools were used:

1. **SHAP (SHapley Additive exPlanations):**
   - Based on cooperative game theory.
   - Quantifies contribution of each feature to the prediction.
   - Provides both **global** and **local** interpretability.
2. **LIME (Local Interpretable Model-Agnostic Explanations):**
   - Locally approximates the model around a specific prediction.
   - Helps explain why an individual patient was classified as diseased or healthy.

Improves model transparency and clinician trust.

SHAP and LIME based Explainability Framework for Disease Diagnosis


Trained Ensemble Model
LIME Explainer
SHAP Explainer
LIME Feature Contribution Estimation
Input Patient Data
SHAP Value Computation
LIME Top-15 Feature Plot
Model Prediction
SHAP Summary Plot
SHAP Feature Importance Plot
Local Interpretability - Why This Prediction Was Made
Global Interpretability - Feature Influence Across All Patients

Performance Results – Accuracy

Disease	Accuracy (%)
Heart Disease	85.2
Diabetes	81.2
CKD	98.8
Parkinson’s	92.3
Breast Cancer	98.2

- Ensemble models outperform single classifiers.
- CKD and Breast Cancer show near-perfect accuracy due to distinct biomarkers.
- Heart and Diabetes models improved significantly with SMOTETomek balancing.

Comparative Discussion

Ensemble models demonstrated **higher accuracy and stability**.

SHAP and LIME improved model **transparency and clinical trust**.

Class balancing techniques enhanced model **sensitivity**.

Results validate that **Explainable Ensemble Learning** is effective for medical prediction tasks.

Framework can be extended to **multi-modal or multi-disease prediction systems**.

Conclusion & Future Work

Conclusion:

- Developed ensemble-based, explainable AI models for five diseases.
- Achieved high accuracy and interpretability.
- XAI ensures transparency in AI-assisted diagnosis.

Future Work:

- Integrate genetic, imaging, and clinical data for multi-modal prediction.
- Develop real-time hospital deployment system.
- Explore deep ensemble and federated learning for privacy-preserving diagnostics.

Thank You