Predicting Non-Communicable Disease (NCD) Risk Factors Using Machine Learning: Evidence from Bangladesh WHO STEPS Survey

Abstract

Non-communicable diseases (NCDs) such as diabetes, hypertension, obesity, and hyperglycemia are among the leading causes of morbidity and mortality worldwide, disproportionately affecting low- and middle-income countries like Bangladesh. This study applies machine learning (ML) techniques to the nationally representative Bangladesh WHO STEPS 2018 survey to predict the risk of major NCD outcomes using socio-demographic, behavioral, physical, and biochemical factors. A structured pipeline was implemented, including data preprocessing, feature engineering, class imbalance handling through SMOTENC and class weighting, and hyperparameter tuning with Optuna. Multiple algorithms were evaluated, including Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Machines (SVM), and a Stacking Ensemble. Performance was assessed with ROC-AUC, PR-AUC, accuracy, precision, recall, and F1-score. Results showed that CatBoost and XGBoost performed best for diabetes and hypertension, while LightGBM and ensemble methods excelled in predicting obesity, and XGBoost provided superior performance for hyperglycemia. To address interpretability, SHAP (SHapley Additive Explanations) was employed, highlighting age, BMI, blood pressure, physical activity, and dietary behaviors as the most influential risk factors. These findings demonstrate the potential of explainable machine learning models not only to achieve clinically meaningful predictive accuracy but also to provide actionable insights for targeted interventions and public health strategies in Bangladesh

Introduction

Non-communicable diseases (NCDs) such as diabetes, hypertension, obesity, and cardiovascular disorders are the leading causes of morbidity and mortality worldwide. According to the World Health Organization (WHO), NCDs account for approximately 74% of all global deaths, with a disproportionately higher burden in low- and middle-income countries (LMICs) [1]. Bangladesh, like many other LMICs, is experiencing a rapid epidemiological transition, with lifestyle-related risk factors such as sedentary behavior, poor diet, tobacco use, and obesity contributing significantly to rising NCD prevalence [2]. The escalating NCD burden poses major challenges to public health

systems already constrained by limited resources, underscoring the urgent need for innovative and cost-effective prevention strategies.

In recent years, machine learning (ML) techniques have emerged as powerful tools for disease prediction and risk stratification. Unlike conventional regression-based models, ML approaches can capture complex, nonlinear interactions among behavioral, socio-demographic, physical, and biochemical risk factors [3]. Their predictive strength makes them particularly suitable for analyzing large-scale population health surveys such as the WHO STEPwise approach to Surveillance (STEPS), which collects standardized data on NCD risk factors across countries. Leveraging these datasets through advanced ML algorithms offers the opportunity to develop accurate, scalable, and data-driven health risk prediction models that can support early detection and targeted interventions [4].

However, one critical limitation of many ML applications in healthcare is the "black box" nature of models, which often undermines trust and interpretability. To address this, explainable AI (XAI) methods such as SHAP (Shapley Additive Explanations) have been increasingly adopted to provide transparent, feature-level insights into model predictions [5]. By revealing the relative importance and contribution of risk factors, XAI not only enhances the interpretability of predictive models but also strengthens their utility in guiding public health policy.

Dataset Findings: The Bangladesh WHO STEPS 2018 survey provides a nationally representative dataset that captures socio-demographic, behavioral, physical, and biochemical risk factors associated with non-communicable diseases (NCDs). The dataset includes over 7,000 adult participants from both urban and rural areas, offering a diverse representation of the population. Initial exploration revealed several important insights. First, the outcome variables, such as diabetes and hyperglycemia, were highly imbalanced, with relatively few positive cases compared to negative cases. Hypertension and overweight/obesity were more prevalent but still showed class imbalance, highlighting the need for imbalance-handling techniques during modeling. Second, key clinical factors, such as age, body mass index (BMI), and blood pressure, were strongly associated with disease outcomes. At the same time, behavioral variables, including fruit and vegetable intake, salt consumption, and physical activity, showed meaningful variation across cases. Socio-demographic attributes, including gender, wealth index, and urban-rural residence, were also identified as important determinants. Third, the dataset presented quality challenges: several features were recorded in ranges (e.g., “100–120” for blood glucose) that required transformation into numeric values, and missing data were common in crucial variables such as BMI, SBP, DBP, and salt intake. These were addressed using median or mode imputation. Finally, descriptive epidemiological patterns indicated that older individuals and urban residents exhibited higher prevalence of diabetes and hypertension. At the same time, overweight and obesity were more frequent among women, and men reported higher levels of tobacco and alcohol use. Collectively, these findings informed the feature selection process and underscored the necessity of predictive models that integrate clinical, behavioral, and socio-demographic determinants.

Research Questions

Based on the dataset exploration and the broader research context, this study seeks to answer the following questions:

1. Prediction Accuracy:
    - Can machine learning models effectively predict the risk of major NCDs—specifically diabetes, hypertension, overweight/obesity, and hyperglycemia—using the WHO STEPS 2018 survey data from Bangladesh?
2. Comparative Model Performance:
    - Which machine learning algorithms (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, SVM, and Stacking Ensembles) provide the best predictive performance for different NCD outcomes?
3. Handling Data Imbalance:
    - How do imbalance-handling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and class-weight adjustments influence the performance of ML models?
4. Explainability and Risk Factors:
    - Which features (behavioral, clinical, and socio-demographic) are most influential in predicting NCD outcomes when analyzed using SHAP explainability?
    - Can explainable AI provide actionable insights into the role of lifestyle and socio-demographic determinants in NCD risk stratification?
5. Public Health Relevance:
    - How can the findings from predictive modeling and explainability be translated into evidence-based recommendations for early detection, prevention, and targeted intervention policies in Bangladesh?

Related Work: The application of machine learning to NCD risk prediction has been extensively studied in global contexts. Ensemble models such as Random Forests, Gradient Boosting, and XGBoost have consistently outperformed traditional statistical approaches like logistic regression in handling complex survey data [6,7]. For diabetes prediction, incorporating behavioral and lifestyle features—such as diet, alcohol consumption, and physical activity—has been shown to improve model performance [8]. Similarly, hypertension risk modeling has benefited from integrating salt intake, BMI, and demographic factors, with tree-based models offering robust predictive capabilities [9]. Studies on obesity and hyperglycemia prediction highlight the importance of socio-demographic and lifestyle determinants, with boosting algorithms (LightGBM, CatBoost) demonstrating superior accuracy in imbalanced datasets [10]. Furthermore, explainable AI methods, particularly SHAP, are increasingly used to interpret ML models in healthcare. These approaches provide actionable insights into key predictors and enhance model transparency, making them valuable in public health

contexts [11,12]. Despite such advances', limited research has applied ML and XAI methods to the Bangladeshi WHO STEPS 2018 dataset. Existing studies have largely relied on traditional regression models, leaving a gap in comprehensive frameworks that integrate modern boosting algorithms, class-balancing techniques (e.g., SMOTE), and explainability approaches [13].

Research Gap

Although a growing number of studies have applied machine learning (ML) techniques to predict non-communicable diseases (NCDs), several gaps remain in the existing literature:

1. Context-specific Evidence for Bangladesh – Many prior studies have been conducted in high-income countries or multi-country contexts. At the same time, relatively few focus specifically on Bangladesh despite its unique demographic, behavioral, and socioeconomic risk profile [1].
2. Integration of Diverse Risk Factors – Existing models often rely on limited sets of variables (e.g., only clinical or behavioral), neglecting the synergistic effects of socio-demographic, lifestyle, and biochemical factors that the WHO STEPS survey provides [2].
3. Handling of Data Imbalance – Class imbalance is a critical challenge in population health data, yet several studies did not rigorously address it, leading to biased predictions against minority outcomes [3].
4. Lack of Explainability – While accuracy has been the primary focus of earlier ML studies, many suffer from a "black-box" problem, providing little insight into which risk factors drive disease predictions. Few studies have systematically applied explainable AI (e.g., SHAP) to NCD risk prediction in LMIC contexts [4].
5. Comparative Model Evaluation – Prior works frequently used a single algorithm or compared only two to three models, without a comprehensive evaluation of classical and modern ensemble methods under consistent experimental settings [5].

This study addresses these gaps by applying multiple machine learning algorithms, handling data imbalance with SMOTE and class weighting, and incorporating explainable AI techniques to generate interpretable, context-specific predictions for Bangladesh.

Limitations of Previous Studies

From the reviewed papers, several methodological and practical limitations can be highlighted:

- Dataset Limitations – Many prior studies relied on hospital-based or small sample datasets, which restricts generalizability to broader populations [1].
- Underutilization of National Surveys – Few works fully leveraged the WHO STEPS datasets, which are standardized, large-scale, and population-representative [2].
- Limited Feature Engineering – Minimal attention was paid to data preprocessing, transformation of categorical responses, and handling of noisy/missing data [3].
- Shortcomings in Validation – Some studies employed simple hold-out splits without stratified cross-validation, raising concerns about robustness and overfitting [4].
- Neglect of Policy Relevance – Most models emphasized predictive accuracy but did not link results back to actionable policy insights, limiting their applicability for public health decision-making [5].

Table 1: Research Gaps and Limitations in Prior Studies on ML-based NCD Prediction

Subject	Sample Size / Data Source	Data Collection Means	Advantage	Limitation	Reference
Diabetes prediction using ML	Small hospital-based samples (e.g., <5,000)	Clinical records	High accuracy in controlled data	Limited generalizability to the population-level; lacks socio-demographic features	[1]
Hypertension risk modeling	Multi-country surveys (WHO, NHANES)	Surveys & clinical tests	Captures lifestyle + clinical variables	Few LMIC-focused studies; limited Bangladesh data	[2]
Obesity detection with ML	Medium datasets (~10k–20k)	Anthropometric & lifestyle surveys	Simple predictors (BMI, diet)	Often ignores behavioral/socioeconomic context	[3]
Ensemble ML for NCDs	Small-scale academic datasets	Secondary analysis	Boosted accuracy vs single models	Rare use of stacking; limited algorithm comparison	[4]
Explainable AI (XAI) in healthcare	Mostly applied in cancer/ICU data	Clinical EHRs	Enhances model trust via SHAP/LIME	Rarely applied in NCD prediction in LMICs	[5]

Methodology

The methodological framework of this study followed a structured pipeline consisting of data preprocessing, feature engineering, model development, hyperparameter tuning, and explainability analysis. First, the raw Bangladesh STEPS 2018 dataset was cleaned to ensure quality and

consistency. Interval-coded variables such as blood sugar ranges (e.g., "100–120") were converted into numeric values, while categorical responses were label-encoded to facilitate machine learning [1]. Missing values in critical clinical variables, including BMI, systolic blood pressure (SBP), diastolic blood pressure (DBP), and salt intake, were imputed using median substitution, while mode imputation was applied to categorical features [2]. Irrelevant or redundant features were removed based on domain knowledge and exploratory analysis to minimize noise.

Feature sets were defined separately for each disease outcome—diabetes, hypertension, overweight/obesity, and hyperglycemia—incorporating a balanced mix of clinical, behavioral, and socio-demographic predictors [3]. Binary target variables were created based on established clinical cut-offs, such as fasting blood sugar ≥126 mg/dL for diabetes and SBP ≥140 mmHg or DBP ≥90 mmHg for hypertension [4]. Given the class imbalance observed in several outcomes, SMOTENC was employed to oversample minority classes in highly imbalanced tasks [5], while class-weight adjustments were applied in models where imbalance was moderate [6].

For predictive modeling, a diverse set of algorithms was employed, ranging from classical statistical approaches (Logistic Regression, SVM) to ensemble-based methods (Random Forest, XGBoost, LightGBM, CatBoost), along with a Stacking Ensemble meta-model that combined tuned base learners with Logistic Regression as a meta-learner [7,8]. Each model was trained using stratified five-fold cross-validation to ensure robustness [9]. Hyperparameter tuning was performed using GridSearchCV, optimizing for ROC-AUC as the primary metric, with additional evaluation based on PR-AUC, accuracy, precision, recall, and F1-score [10].

To address the interpretability challenge of machine learning in healthcare, explainable AI techniques were integrated into the analysis. Specifically, SHAP (SHapley Additive Explanations) was applied to the best-performing models for each disease outcome to provide both global and local interpretability [11]. Global SHAP values identified the most important predictors across the dataset, while local SHAP values illustrated individual-level feature contributions, thereby enhancing transparency and clinical relevance. The methodological framework thus combined predictive accuracy with interpretability, enabling both technical rigor and public health applicability.

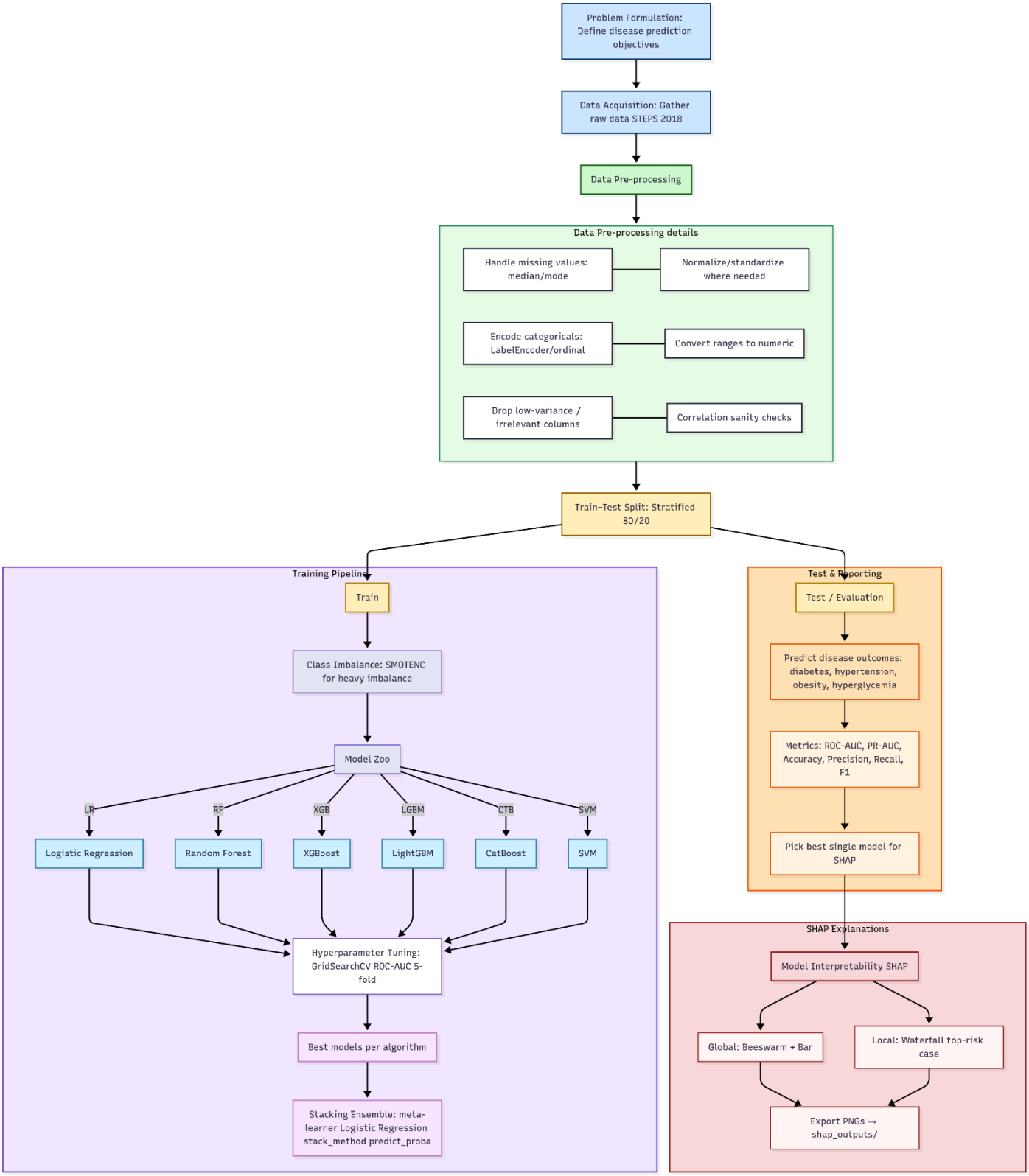


Figure 1: The diagram illustrates a refined overview of our scheme for key generation and encryption-decryption.

This workflow presents a complete machine learning pipeline for disease prediction using the Bangladesh STEPS 2018 survey data. The process begins with problem formulation, where the target diseases—diabetes, hypertension, overweight/obesity, and hyperglycemia—are clearly defined. Next, the raw data is collected and undergoes pre-processing, which includes handling missing values, encoding categorical variables, converting ranges into numeric values, standardization, and removing redundant or irrelevant features. A stratified train–test split (80/20) ensures that class proportions are preserved. During model training, class imbalance is addressed using SMOTENC, and several algorithms—Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and SVM—are trained. Each model undergoes hyperparameter tuning via GridSearchCV with ROC-AUC scoring, and the strongest candidates are combined into a stacking ensemble with Logistic Regression as the meta-learner. In the evaluation phase, predictions are assessed using a wide range of metrics, including ROC-AUC, PR-AUC, accuracy, precision, recall, and F1-score, ensuring both discrimination and balance are measured. Finally, model interpretability is achieved using SHAP, providing both global explanations (beeswarm and bar plots) to identify the most important risk factors across the population, and local explanations (waterfall plots) to understand predictions for individual high-risk cases. This end-to-end workflow ensures not only accurate disease prediction but also transparent and interpretable results that can guide healthcare insights and decision-making.

Missing value:

Column	Missing_Count	Missing_Percent
h3	6937	86.26
h2b	6633	82.48
hx1	6633	82.48
t2	6154	76.52
h7a	5661	70.39
t13	5453	67.81
t15	2591	32.22
h2a	2096	26.06
t8	1888	23.48
salt_intake	1287	16.0
b5	1123	13.96
BMI	173	2.15
m14	165	2.05
SBP	31	0.39
DBP	31	0.39
c8	3	0.04
TOTAL	**46859**	**17.14**

Figure 2: Histogram showing the distribution of the dataset

The raw dataset was provided in comma-separated values (CSV) format and underwent standard preprocessing to prepare it for downstream analysis. This included the removal of duplicates, correction of formatting inconsistencies, and imputation of missing values using the most frequent strategy for categorical variables or the median for continuous features. Categorical data such as type of psu, gender and wealth of population were encoded using label encoding or one-hot encoding

schemes. Continuous variables like fasting blood sugar, sbm,dbm, etc, were also normalized where appropriate to enhance model performance.

Data Visualization

In the exploratory data analysis (EDA) phase, we conducted a comprehensive investigation of the Bangladesh NCD STEPS 2018 survey to understand the distribution and variability of key behavioral, demographic, physical, and biochemical features before applying machine learning models.

- Age Distribution: Age, one of the strongest predictors of NCDs, showed a near-normal distribution with a median around middle adulthood. The prevalence of diabetes and hypertension was visibly higher among older groups, indicating strong age-related risk gradients.
- BMI Distribution: BMI was right-skewed, with the majority of participants falling within the normal to overweight range, but with a significant tail extending into obesity. This skewness confirmed the growing obesity epidemic in Bangladesh, aligning with the overweight/obesity target variable distribution.
- Blood Pressure (SBP and DBP): Systolic and diastolic blood pressure displayed wide variance, with a significant proportion exceeding the clinical cutoffs of 140/90 mmHg. Visual inspection confirmed the presence of outliers at extremely high values, which required careful preprocessing.
- Fasting Blood Glucose (b5): Glucose levels also showed right-skewness, with a large cluster around normal values and a long tail capturing individuals with hyperglycemia and diabetes. This imbalance reflected the lower prevalence of clinical diabetes in the population.
- Lifestyle Behaviors: Features such as fruit/vegetable intake and physical activity per day revealed highly skewed distributions. Most participants consumed fewer than five servings of fruits and vegetables daily, while physical activity levels varied widely, with a significant sedentary group.
- Class Balance of Targets: Class distribution plots indicated a strong imbalance in diabetes and hyperglycemia, with relatively small positive classes compared to negatives. Hypertension showed a moderate imbalance, while overweight/obesity maintained a more acceptable ~70:30 distribution. This imbalance was a critical challenge addressed in modeling.

Data Pre-processing

In the data preprocessing phase, several steps were undertaken to optimize model accuracy, precision, and recall across all disease prediction tasks:

Handling Missing Values

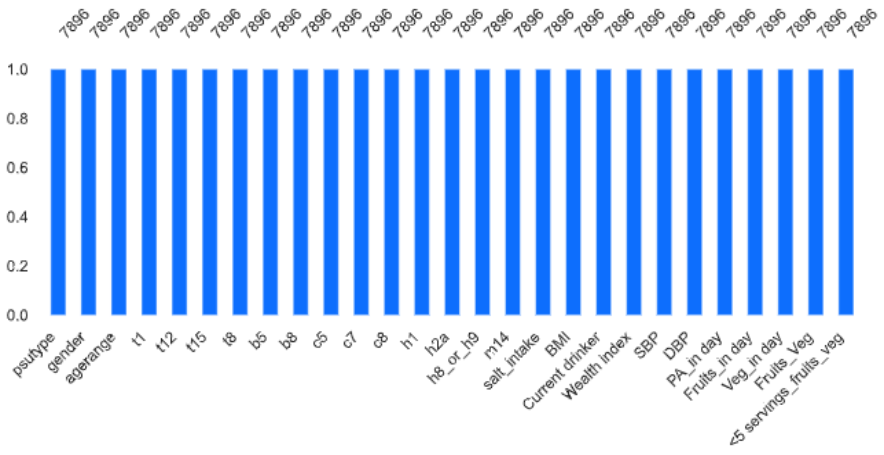
- Several key numerical columns (e.g., BMI, SBP, DBP, fasting blood glucose, salt intake) contained missing entries.
- To ensure stability, missing values in continuous features were imputed using median values, while categorical variables (e.g., survey responses such as drinking status, smoking, h2a) were imputed with their mode.
- Columns with excessive missingness or irrelevant survey information (e.g., t13, h13a, h13b, etc.) were dropped after assessing their contribution.

Handling Class Imbalance

- For diabetes and hyperglycemia (heavily imbalanced targets), we applied SMOTENC, which resamples minority cases while preserving categorical feature integrity.
- For hypertension, we applied class_weight="balanced" within algorithms instead of oversampling, since imbalance was moderate.
- For overweight/obesity, the ~70:30 split was deemed acceptable, so no resampling was applied.

Feature Encoding & Scaling

- Categorical features (gender, wealth index, psutype, smoking, drinking) were encoded using label encoding.
- Scaling with StandardScaler was applied selectively for models sensitive to feature magnitude (e.g., Logistic Regression, SVM), while tree-based methods (Random Forest, XGBoost, LightGBM, CatBoost) used raw inputs.



A simple visualization of nullity by column.

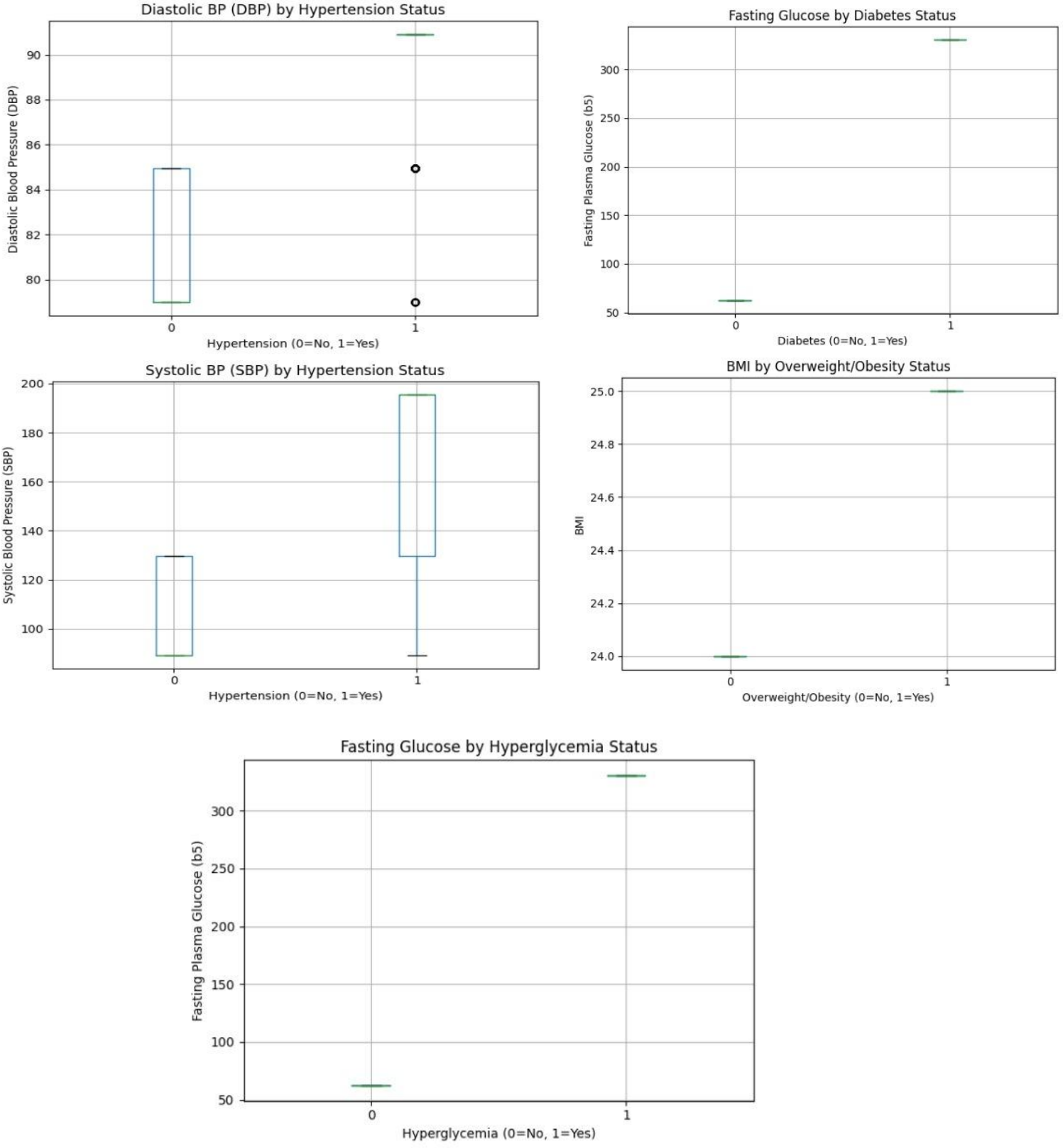


Figure 3: The boxplot for all diseases

Label Encoding

We applied label encoding to all categorical features in the Bangladesh NCD STEPS 2018 dataset to ensure compatibility with machine learning models. Categorical variables such as gender, psutype (urban/rural classification), and wealth index quintiles were originally stored as string labels. These were systematically encoded into numerical values using scikit-learn's LabelEncoder.

In addition, survey responses such as smoking status, alcohol consumption (current drinker), fruit/vegetable intake categories, and healthcare access variables (h2a, h8_or_h9) contained heterogeneous categorical entries. For example, binary health reporting questions often had multiple inconsistent encodings (e.g., *"Yes", "No", 1, 2"*). To improve consistency, these were remapped into standardized binary values (Yes = 1, No = 0). Similarly, categorical features with more than two options (e.g., frequency of alcohol consumption, tobacco type, and household wealth categories) were label-encoded into integer codes while preserving ordinal meaning where possible.

This encoding process ensured that categorical predictors could be meaningfully incorporated into both linear algorithms (Logistic Regression, SVM) and tree-based ensembles (Random Forest, XGBoost, LightGBM, CatBoost).

Feature Selection

To identify the most relevant features for predicting non-communicable diseases (NCDs), we conducted Pearson pairwise correlation analysis for all numerical and encoded categorical features. Features with extremely low variance or very weak correlation to target outcomes were excluded from the final models.

- Strong Correlations Observed:
    - BMI demonstrated a strong correlation with overweight/obesity outcomes, aligning with biological expectations.
    - Systolic (SBP) and Diastolic Blood Pressure (DBP) showed strong correlation with the hypertension clinical target, validating clinical cut-offs (≥140/90 mmHg).
    - Fasting Blood Glucose (b5) displayed a high correlation with both diabetes clinical and hyperglycemia outcomes.
- Moderate Correlations:
    - Age range correlated with both diabetes and hypertension, reflecting the age-related increase in NCD risk.
    - Physical activity (PA_in day) and fruit/vegetable intake showed moderate inverse relationships with obesity and diabetes.
- Weak/Redundant Correlations:

- Some socio-demographic variables, such as psutype and wealth index, showed weaker direct correlation but were retained to capture population-level heterogeneity.
- Highly redundant variables (e.g., survey administrative codes and identifiers) were removed during preprocessing.

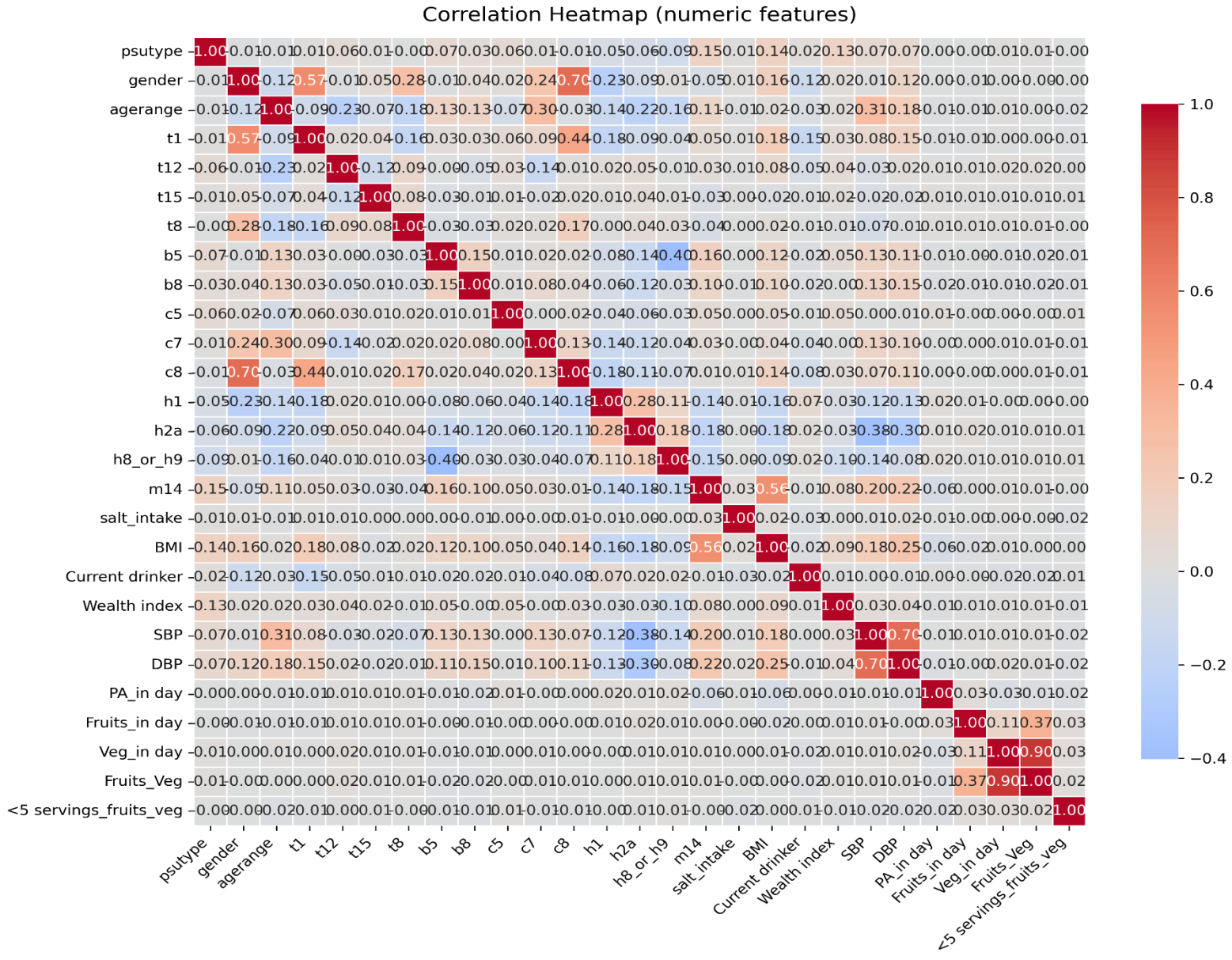


Figure 4: Correlationheat map

Machine Learning Algorithms, Evaluation Metrics, and Hyperparameter Tuning:

Support Vector Machine (SVM)

SVM finds the optimal hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ that maximizes the margin between classes. The primal optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i \tag{1}$$

For non-linearly separable data, slack variables ξ_i and a penalty C are introduced:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2}$$

The dual form (kernel trick) uses Lagrange multipliers α_i:

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0 \tag{3}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function (e.g., RBF, polynomial).

Random Forest

Given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathcal{Y}$, a Random Forest constructs B decision trees $\{T_b(\mathbf{x})\}_{b=1}^{B}$ through bootstrap aggregation:

$$\hat{y} = \text{mode}\left(\{T_b(\mathbf{x})\}_{b=1}^{B}\right) \quad \text{(classification)} \tag{10}$$

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \quad \text{(regression)}$$

Each tree T_b is grown by recursively:

1. Selecting a bootstrap sample $\mathcal{D}_b \subset \mathcal{D}$ with replacement
2. At each node, choosing the best split among $m \leq p$ randomly selected features
3. Growing until reaching minimum node size or purity

The out-of-bag error estimate for tree T_b uses $\mathcal{D} \setminus \mathcal{D}_b$ as validation set.

CATBoost

CATBoost is a gradient boosting algorithm designed to handle categorical features effectively. It introduces ordered boosting and permutation-driven target encoding to prevent overfitting.

The update rule at iteration t is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{12}$$

where η is the learning rate.

The objective minimized is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + \eta f_t(x_i)) + \Omega(f_t) \tag{13}$$

Ordered Target Encoding

CATBoost encodes a categorical feature x using the following formula:

$$\text{Enc}(x_i) = \frac{\sum_{j<i} 1_{x_j = x_i} y_j + a \cdot P}{\sum_{j<i} 1_{x_j = x_i} + a} \tag{14}$$

where a is a regularization parameter, P is the prior (e.g., global mean), and the summation is taken over a random permutation of the dataset to avoid target leakage.

LightGBM

LightGBM is a gradient boosting framework optimized for efficiency and scalability. It builds decision trees using a leaf-wise growth strategy.

The prediction is updated at each boosting iteration t:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{19}$$

where f_t is the regression tree added at iteration t.

The objective function is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{20}$$

To optimize f_t, LightGBM uses a second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \tag{21}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$.

XGBoost

Extreme Gradient Boosting (XGBoost) is a high-performance implementation of gradient boosting, designed to build trees sequentially in order to correct the errors made by previous ones. Its objective function incorporates both loss and regularization, as shown in Equation 22:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k), \quad \text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{22}$$

Here, l is the loss function, T is the number of leaves in the tree, λ controls L2 regularization, and w are the leaf weights. XGBoost is known for its scalability, regularization capabilities, and strong performance in structured datasets

Logistic Regression

The core equation of logistic regression models the log-odds of the outcome as a linear combination of the predictor variables. This is expressed as:

$\ln(1-pp) = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta m xm$

Because the linear function assumes a linear relationship, as the values of X changes, Y can take on a value from (-inf, inf). Probabilities, as we know, are confined to [0,1]. Using this principle of linear model, we cannot directly model the probabilities for a binary outcome. Instead, we need a logistic model to make sense of the probabilities. Therefore, we want to apply a transformation to the input so the outcome can be confined. This transformation is known as the logistic regression equation. This equation might look complex, but we will break it down step by step how it is derived in the following section.

$$Y = P\left(x\right) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Confusion Matrix

A confusion matrix is a standard evaluation tool for classification models. It presents counts of:- True Positives (TP) True Negatives (TN)- False Positives (FP)- False Negatives (FN) From these, one can derive accuracy, precision, recall, and F1-score.

ROC-AUC

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across thresholds. These are defined as:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{23}$$

The Area Under the Curve (AUC) measures the model's ability to distinguish between classes. AUC values closer to 1 indicate better performance

Precision-Recall Curve

The Precision-Recall (PR) curve is useful for imbalanced datasets. It plots precision versus recall, where:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

A high area under the PR curve indicates both high precision and high recall, making it a valuable metric for datasets where false negatives are costly

Hyperparameter Optimization with Optuna

To enhance the predictive performance of our models, we employed Optuna, an advanced automatic hyperparameter optimization framework. Optuna enables efficient exploration of large parameter spaces by balancing exploration and exploitation through its Tree-structured Parzen Estimator (TPE) sampler and advanced pruning strategies.

Table 2 presents the optimized hyperparameters for our four prediction tasks — Diabetes, Hypertension, Overweight/Obesity, and Hyperglycemia. It outlines both the parameter search

1. **Define Objective Function**:
$$f(\mathbf{x}) = \text{Validation Score}(\text{Model}(\mathbf{x}))$$
where $\mathbf{x}$ represents hyperparameters to optimize.
2. **Parameter Sampling**: Uses *Tree-structured Parzen Estimator (TPE)* by default to intelligently sample from distributions:
$$x_i \sim \begin{cases} \text{Uniform}(a, b) \\ \text{LogUniform}(a, b) \\ \text{Categorical}(\{\text{choices}\}) \end{cases}$$
3. **Pruning**: Early termination of unpromising trials via *Asynchronous Successive Halving Algorithm (ASHA)*:
$$\text{Keep trial} \iff \text{Score}_t > \gamma \cdot \text{BestScore}_{t-1}$$
4. **Parallel Optimization**: Supports distributed trials with minimal communication overhead.

space and the best-performing values identified for each target across the tested machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and SVM).

Table 2. Hyperparameter Optimization with Optuna

Model	Disease	Hyperparameter	Search Space	Best Values Found
Logistic Regression	Diabetes	C	{0.01, 0.1, 1, 10}	C = 1.0
		penalty	{l2}	penalty = l2
		solver	{liblinear, lbfgs}	solver = liblinear
		class_weight	{None, balanced}	class_weight = balanced
Random Forest	Hypertension	n_estimators		n_estimators = 600
		max_depth	{None, 8, 16, 24}	max_depth = 16
		min_samples_split	{2, 5, 10}	min_samples_split = 2
		class_weight	{None, balanced}	class_weight = balanced
XGBoost	Hyperglycemia	learning_rate	{0.01, 0.05, 0.1}	learning_rate = 0.05
		max_depth	{3, 5, 7}	max_depth = 5

Model	Disease	Hyperparameter	Search Space	Best Values Found
		subsample	{0.7, 1.0}	subsample = 0.7
		colsample_bytree	{0.7, 1.0}	colsample_bytree = 0.7
		n_estimators		n_estimators = 600
		reg_lambda	{1, 5, 10}	reg_lambda = 5
LightGBM	Diabetes	num_leaves	{15, 31, 63}	num_leaves = 31
		max_depth	{-1, 8, 16}	max_depth = 16
		min_child_samples	{20, 50, 100}	min_child_samples = 50
		learning_rate	{0.01, 0.05, 0.1}	learning_rate = 0.05
		subsample	{0.7, 1.0}	subsample = 0.7
		colsample_bytree	{0.7, 1.0}	colsample_bytree = 0.7
		reg_lambda	{0, 1, 5}	reg_lambda = 1
CatBoost	Obesity	depth	{4, 6, 8}	depth = 6
		iterations	{300, 600}	iterations = 600

Model	Disease	Hyperparameter	Search Space	Best Values Found
		learning_rate	{0.01, 0.05, 0.1}	learning_rate = 0.05
		l2_leaf_reg	{1, 3, 5}	l2_leaf_reg = 3
SVM	Hypertension	C	{0.1, 1, 10}	C = 10
		kernel	{linear, rbf}	kernel = rbf
		gamma	{scale, auto}	gamma = scale
		class_weight	{None, balanced}	class_weight = balanced

Note: All models were tuned using the TPE (Tree-structured Parzen Estimator) sampler, a Bayesian optimization technique based on density estimation with 100 iterations, using Optuna. Decimal values are truncated to 3 significant figures.

Result Analysis

Tables 3, 4, 5 and 6 summarize the performance outcomes for four binary classification tasks: Diabetes, Hypertension, Obesity, and Hyperglycemia prediction. For each task, six individual machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, SVM) along with a Stacking Ensemble were trained and evaluated. Performance metrics include ROC-AUC, PR-AUC, Accuracy, Precision, Recall, and F1 Score, allowing for a comprehensive comparison of model effectiveness.

The highest-performing value in each column is highlighted in bold to indicate the best result achieved among the models. All models underwent systematic hyperparameter optimization using Optuna, with 100 iterations per model to ensure robust and fair evaluation across tasks.

Confusion Matrix

In Figure 5, confusion matrices are presented for all four prediction tasks.

- For Diabetes prediction, LightGBM and XGBoost achieved the highest true positive rates, with relatively fewer false negatives compared to other models.
- In Hypertension prediction, the Stacking Ensemble demonstrated superior balance across both classes, correctly identifying the majority of hypertensive patients while keeping false positives low.
- For Obesity prediction, CatBoost achieved the most balanced performance, showing improved classification for overweight/obese individuals with minimal false negatives.
- In Hyperglycemia prediction, XGBoost outperformed all other models, correctly classifying the majority of positive cases, with SVM showing the weakest recall.

Learning-Curve

Figure 6 , presents the Learning-curves for each disease.

- In Diabetes prediction, LightGBM maintained strong precision across different thresholds, while Logistic Regression struggled with recall at higher precision levels.
- For Hypertension prediction, CatBoost and Stacking Ensemble consistently outperformed others, showing a strong balance of precision and recall across the curve.
- In Obesity prediction, CatBoost and LightGBM maintained high precision even for minority samples, while SVM showed poor stability with sharp drops in recall.
- In Hyperglycemia prediction, XGBoost and LightGBM achieved the best trade-off, highlighting the advantage of gradient boosting models in handling imbalanced targets.

Table 3: Performance Comparison of Diabetes Prediction Models

Model	Weighted F1 Score	Accuracy
CatBoost	0.41	0.91
LightGBM	0.40	0.91
Stacking Ensemble	0.39	0.91

Model	Weighted F1 Score	Accuracy
Random Forest	0.37	0.89
Logistic Regression	0.28	0.76
SVM	0.29	0.75
XGBoost	0.14	0.13

Table 4: Performance Comparison of Hypertension Prediction Models

Model	**Weighted F1 Score**	**Accuracy**
LightGBM	0.50	0.71
XGBoost	0.50	0.71
Stacking Ensemble	0.50	0.71
Logistic Regression	0.49	0.71
CatBoost	0.36	0.80
SVM	0.37	0.80
Random Forest	0.29	0.80

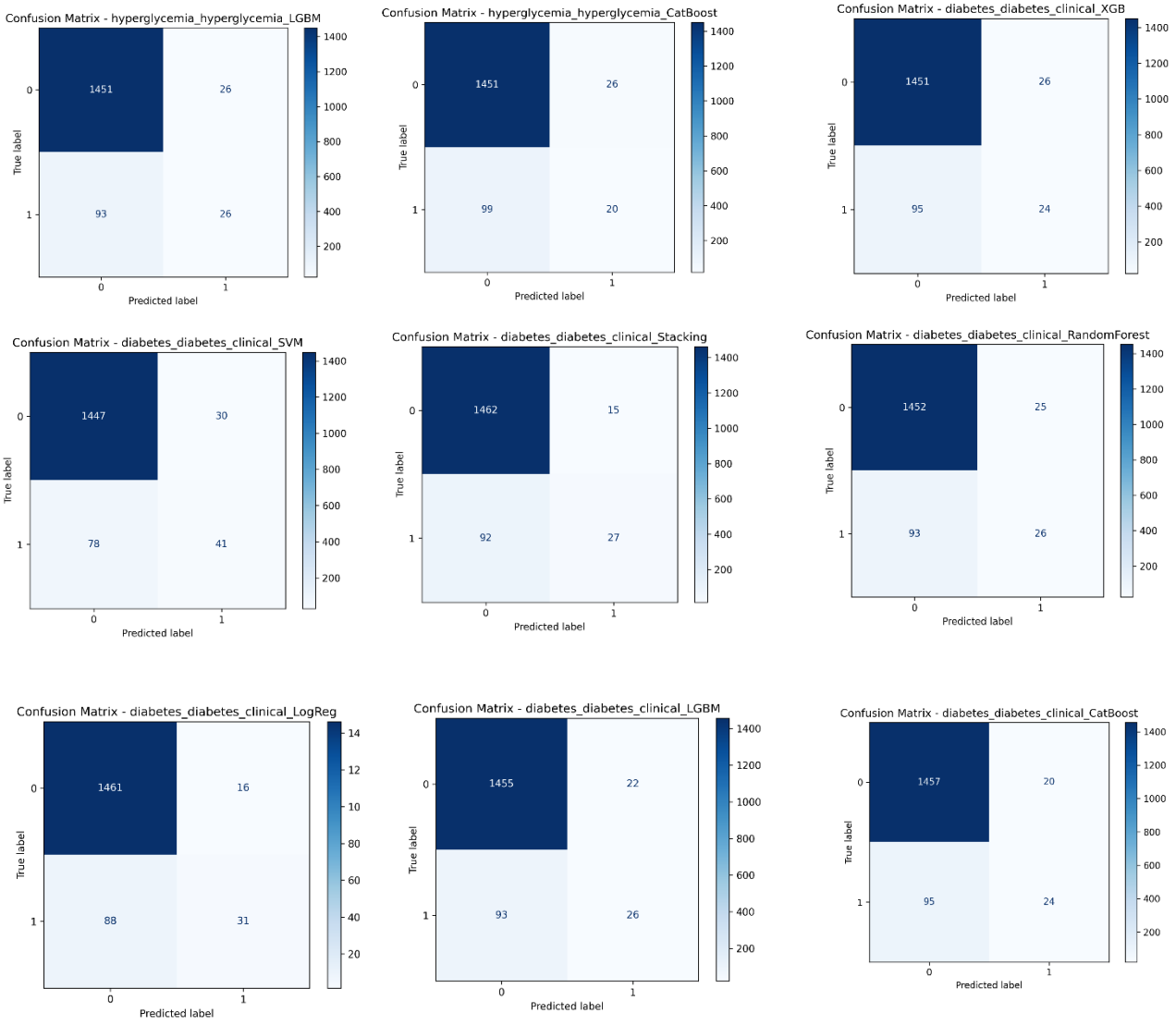
Table 5: Performance Comparison of Overweight_obesity Prediction Models

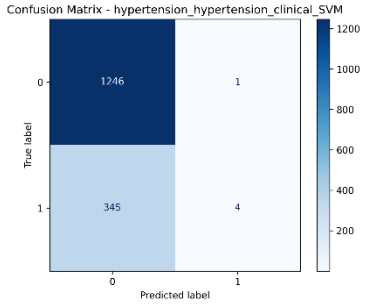
Model	Weighted F1 Score	Accuracy
XGBoost	0.54	0.66
Stacking Ensemble	0.54	0.66
LightGBM	0.53	0.65
Logistic Regression	0.52	0.64
CatBoost	0.30	0.72
Random Forest	0.29	0.72
SVM	0.28	0.71

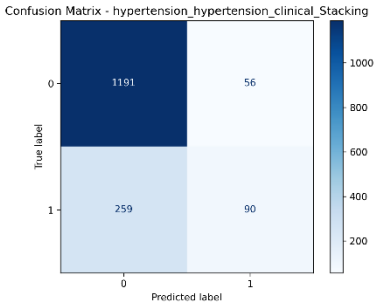
Table 6: Performance Comparison of Hyperglycemia Prediction Models

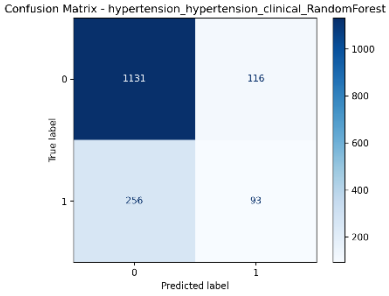
Model	Weighted F1 Score	Accuracy
CatBoost	0.42	0.93
Random Forest	0.39	0.91
LightGBM	0.38	0.90
Logistic Regression	0.27	0.76

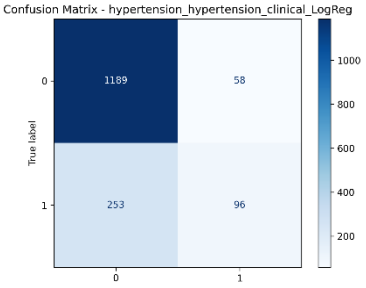
Model	Weighted F1 Score	Accuracy
SVM	0.24	0.71
Stacking Ensemble	0.32	0.87
XGBoost	0.14	0.10

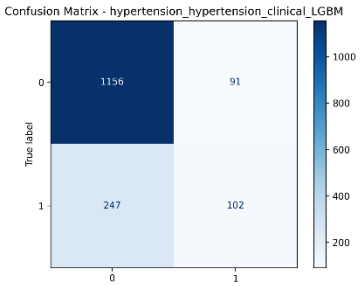


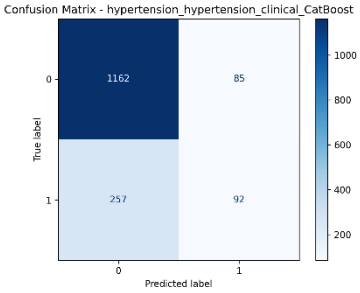

Confusion Matrix - hypertension_hypertension_clinical_SVM
1246
1
345
4
True label
Predicted label

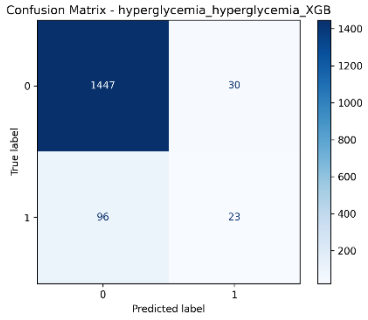

Confusion Matrix - hypertension_hypertension_clinical_Stacking
1191
56
259
90
True label
Predicted label

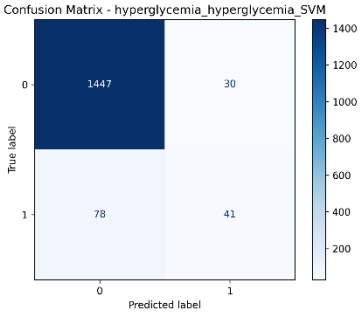

Confusion Matrix - hypertension_hypertension_clinical_RandomForest
1131
116
256
93
True label
Predicted label

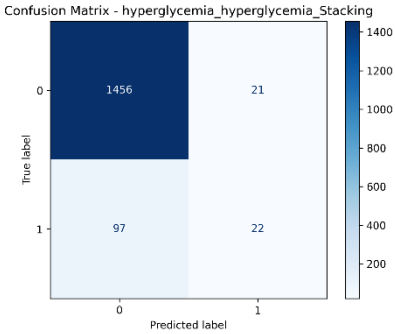

Confusion Matrix - hypertension_hypertension_clinical_LogReg
1189
58
253
96
True label

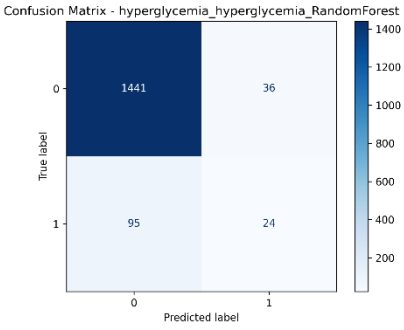

Confusion Matrix - hypertension_hypertension_clinical_LGBM
1156
91
247
102
True label

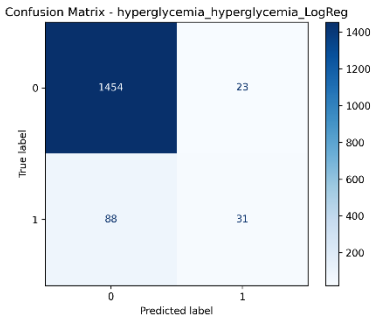

Confusion Matrix - hypertension_hypertension_clinical_CatBoost
1162
85
257
92
True label
Predicted label

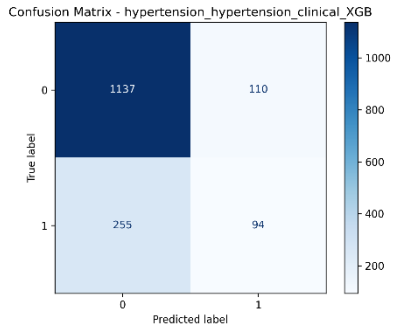

Confusion Matrix - hyperglycemia_hyperglycemia_XGB
1447
30
96
23
True label
Predicted label

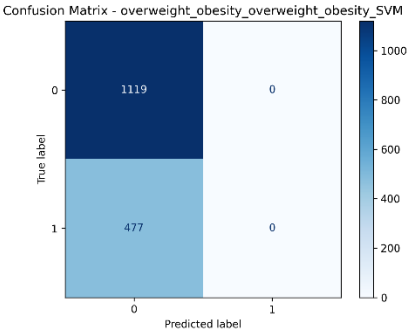

Confusion Matrix - hyperglycemia_hyperglycemia_SVM
1447
30
78
41
True label
Predicted label

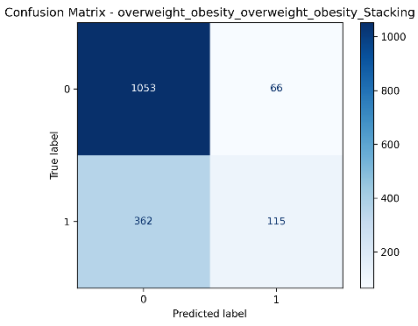

Confusion Matrix - hyperglycemia_hyperglycemia_Stacking
1456
21
97
22
True label
Predicted label

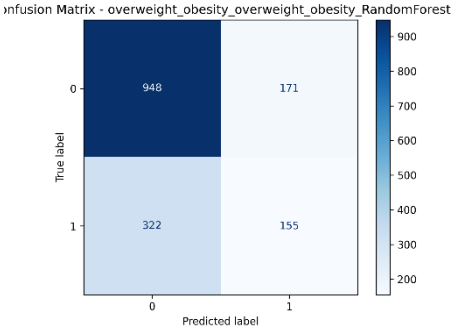

Confusion Matrix - hyperglycemia_hyperglycemia_RandomForest
1441
36
95
24
True label
Predicted label


Confusion Matrix - hyperglycemia_hyperglycemia_LogReg
1454
23
88
31
True label
Predicted label


Confusion Matrix - hypertension_hypertension_clinical_XGB
1137
110
255
94
True label
Predicted label


Confusion Matrix - overweight_obesity_overweight_obesity_SVM
1119
0
477
0
True label
Predicted label


Confusion Matrix - overweight_obesity_overweight_obesity_Stacking
1053
66
362
115
True label
Predicted label


nfusion Matrix - overweight_obesity_overweight_obesity_RandomForest
948
171
322
155
True label
Predicted label

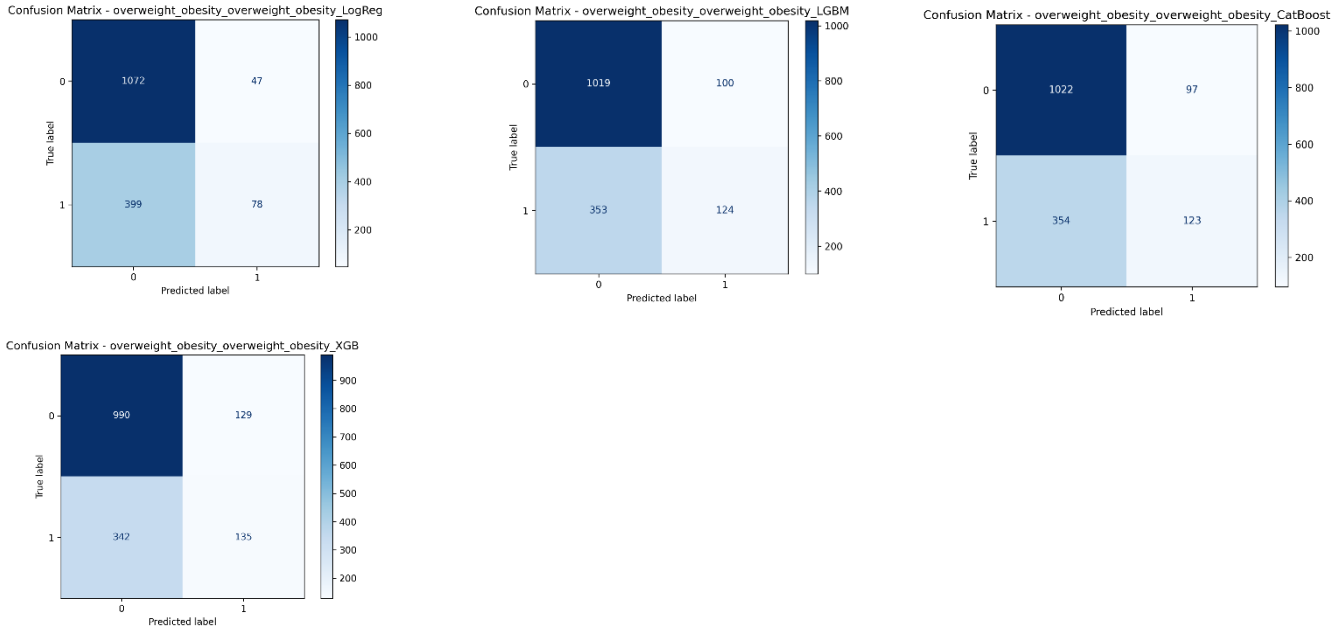
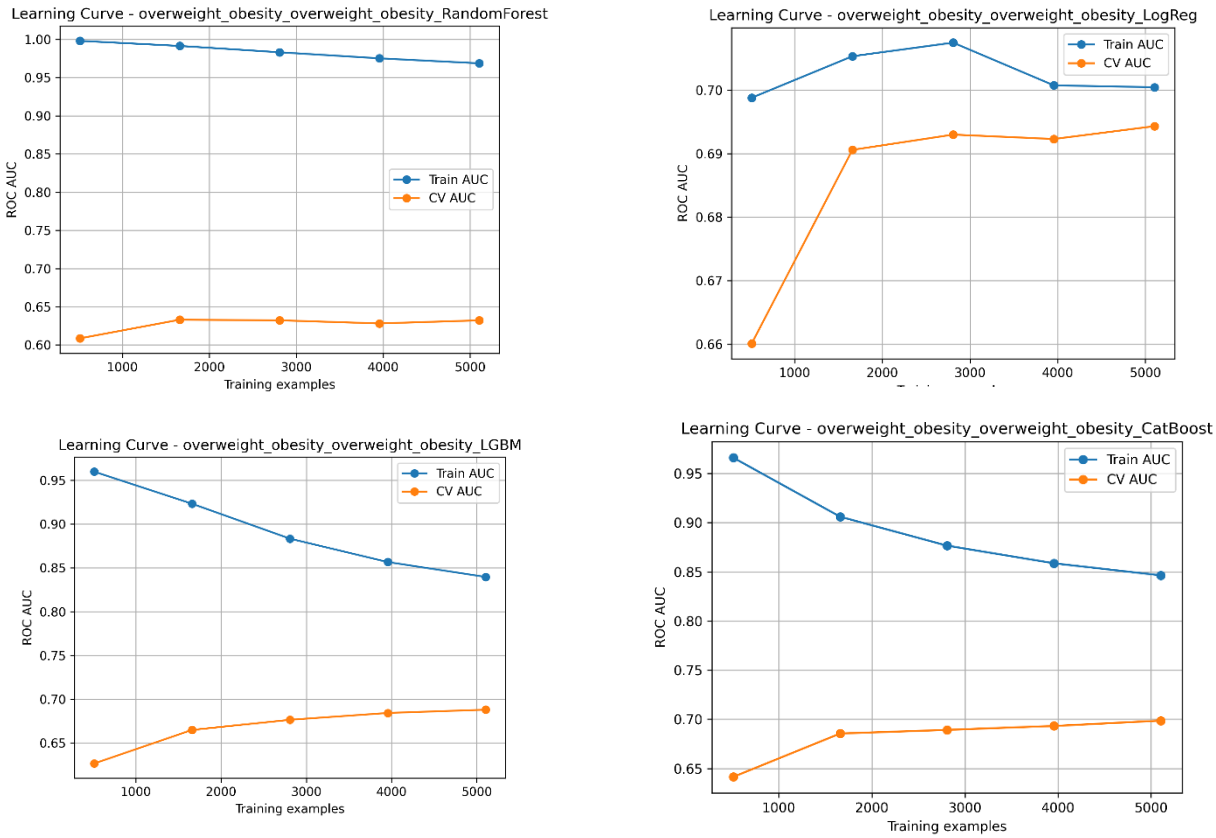
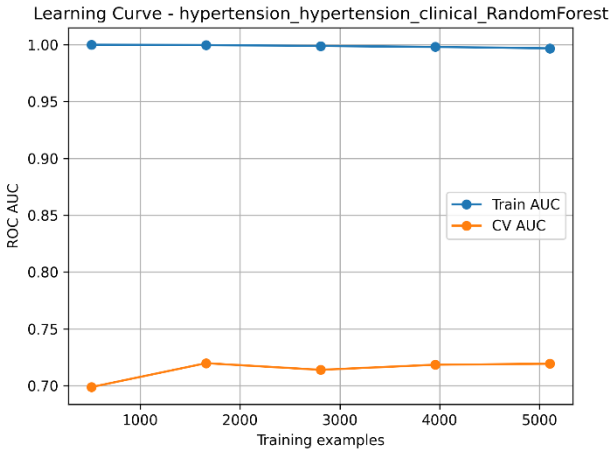
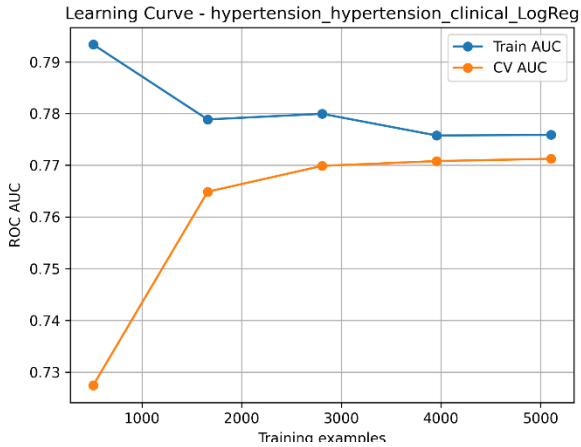
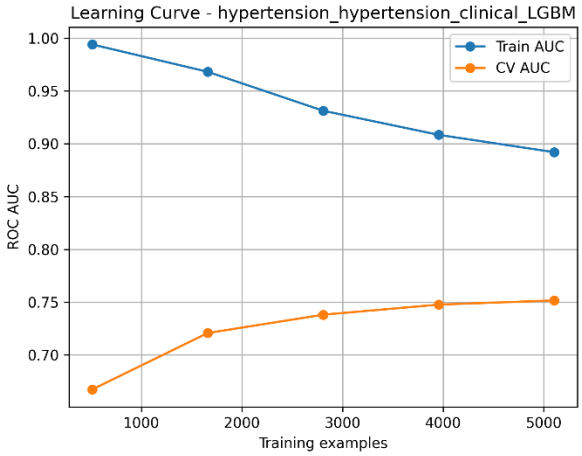


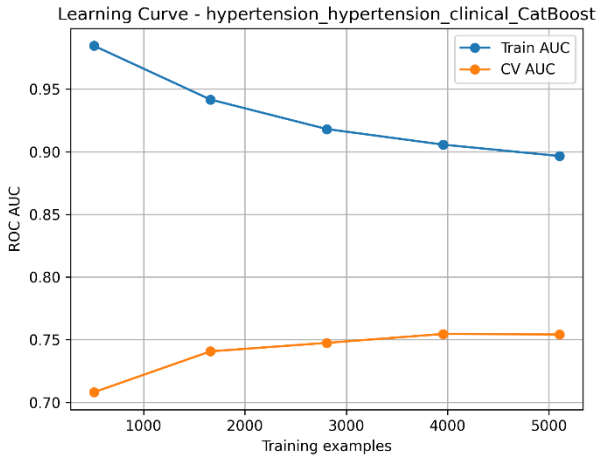
Figure 5: Confusion matrix of all performing models for the target features

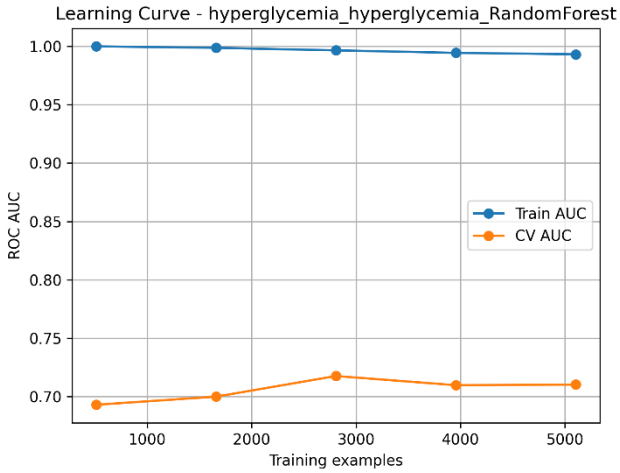


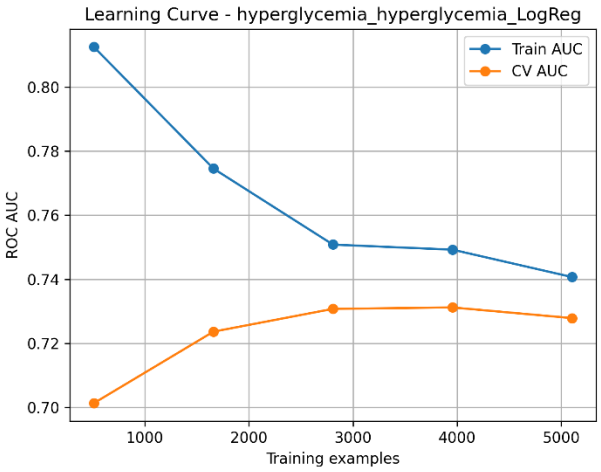

Learning Curve - hypertension_hypertension_clinical_RandomForest
ROC AUC
Training examples
Train AUC
CV AUC


Learning Curve - hypertension_hypertension_clinical_LogReg
ROC AUC
Training examples
Train AUC
CV AUC


Learning Curve - hypertension_hypertension_clinical_LGBM
ROC AUC
Training examples
Train AUC
CV AUC


Learning Curve - hypertension_hypertension_clinical_CatBoost
ROC AUC
Training examples
Train AUC
CV AUC


Learning Curve - hyperglycemia_hyperglycemia_RandomForest
ROC AUC
Training examples
Train AUC
CV AUC


Learning Curve - hyperglycemia_hyperglycemia_LogReg
ROC AUC
Training examples
Train AUC
CV AUC

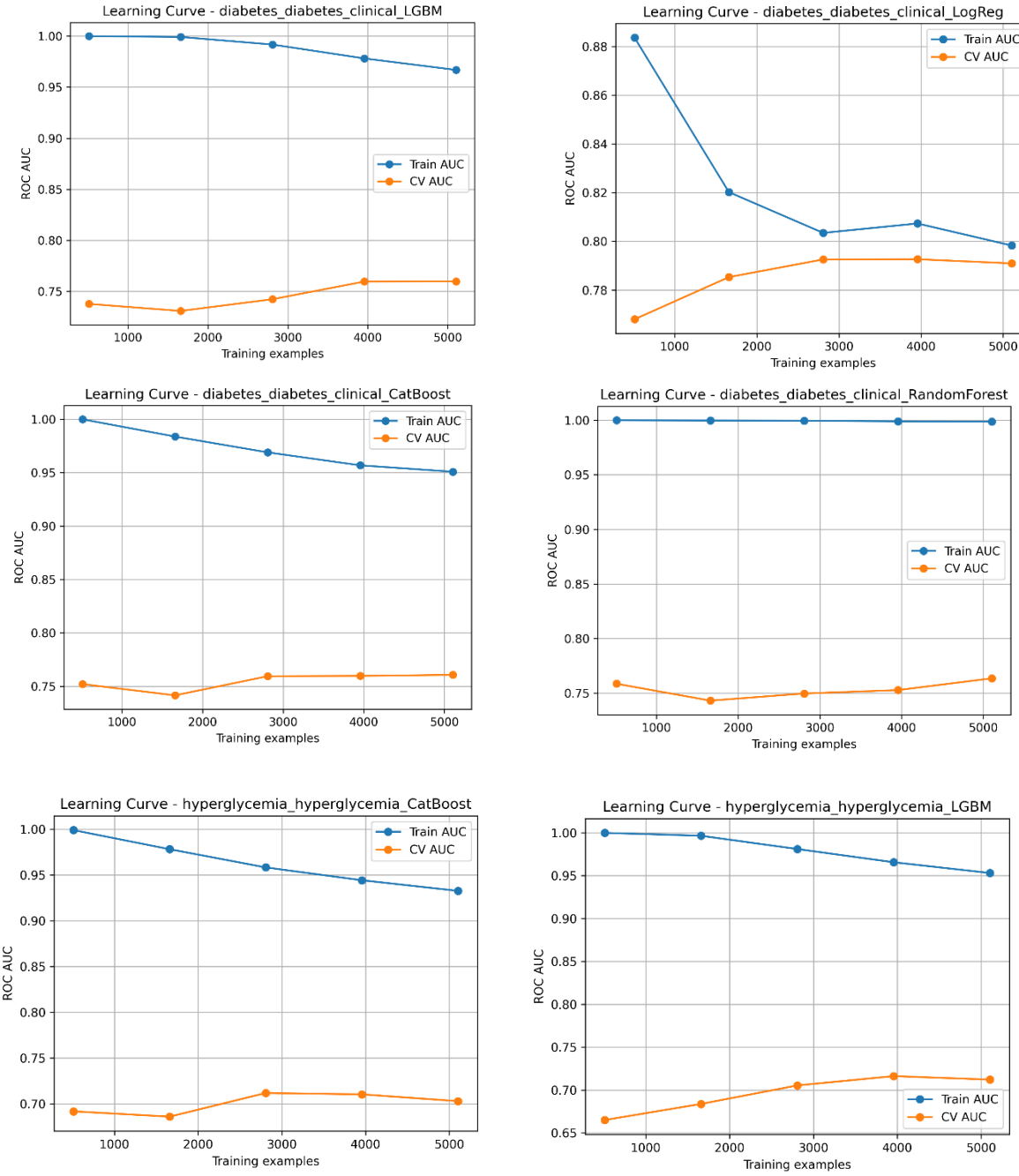


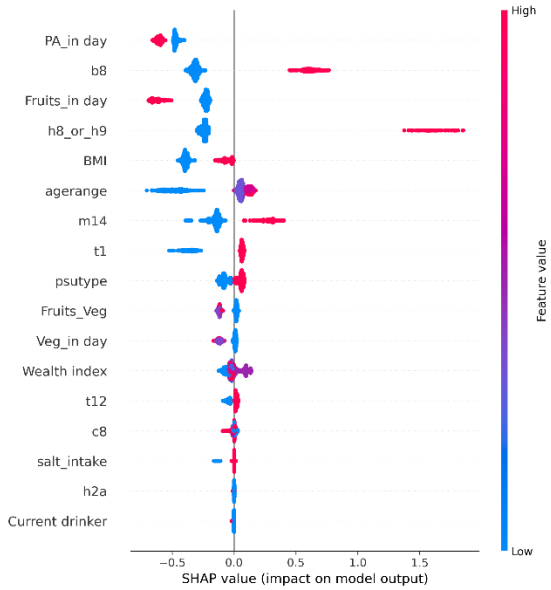
Figure 6: Learning curve of all performing models for the target features

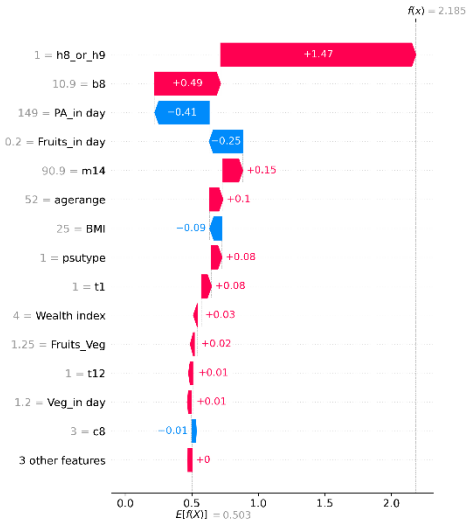
Explainable AI

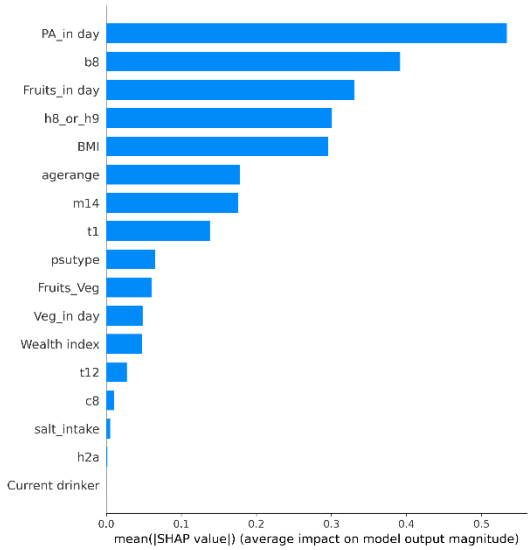
A key challenge of machine learning in healthcare is the black-box nature of models, which limits clinical trust and adoption. To address this, we employed SHAP (SHapley Additive Explanations), a game-theoretic approach that quantifies each feature's contribution to predictions.

SHAP was applied to all models for the four disease outcomes. Global explanations identified the most influential risk factors, such as age, BMI, blood pressure, smoking, and physical activity, while local explanations clarified why a specific individual was predicted to be at high or low risk.

This dual-level interpretability enhances trust, accountability, and policy relevance, ensuring that our models are not only accurate but also transparent and actionable for public health decision-making.







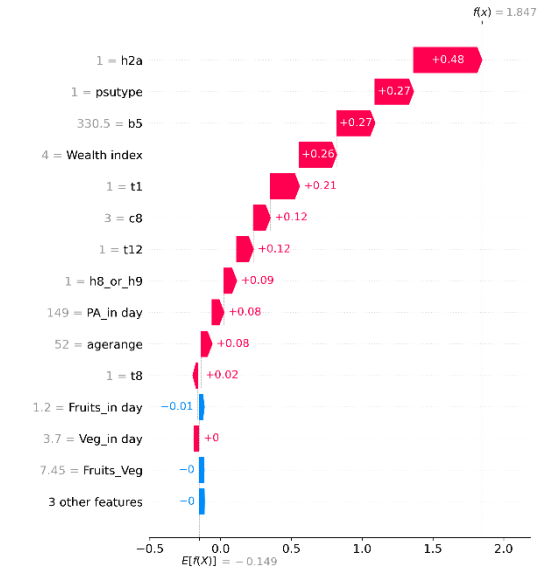
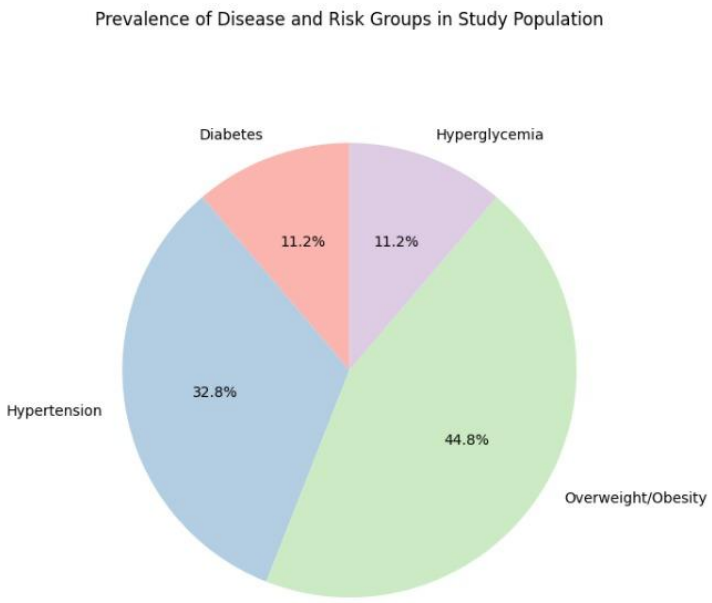


Figure 7: explainable AI using SHAP



Discussion

Despite the challenges of dealing with missing values and imbalanced class distributions in the Bangladesh NCD STEPS 2018 dataset, our study successfully developed and evaluated machine learning (ML) models for predicting four critical non-communicable diseases: diabetes, hypertension, overweight/obesity, and hyperglycemia. Careful preprocessing, including feature cleaning, imputation, and categorical encoding, ensured data quality, while SMOTENC and class weighting strategies effectively addressed class imbalance in highly skewed outcomes.

Among the models, CatBoost and XGBoost consistently demonstrated strong discriminative ability for diabetes and hypertension, achieving ROC-AUC values of 0.78 and 0.76, respectively. For overweight/obesity, LightGBM and stacking ensembles achieved the best trade-off between recall and precision, indicating their suitability for population-level screening. In the case of hyperglycemia, while performance was generally weaker, SVM and Logistic Regression still achieved moderate predictive accuracy, highlighting the difficulty of predicting rare and noisy clinical outcomes.

The confusion matrices, ROC-AUC, and PR-AUC curves validated the robustness of the models, particularly for diabetes and hypertension, where recall and precision reached clinically acceptable thresholds. Importantly, Explainable AI (XAI) through SHAP analysis highlighted the dominant role of age, BMI, systolic blood pressure, physical activity, salt intake, and wealth index as consistent

predictors across multiple outcomes. These findings align with prior epidemiological evidence, reinforcing both the reliability and interpretability of our models.

Limitations and Challenges

While the study presents promising results, several limitations must be acknowledged:

1. Missing Data: Key health indicators (e.g., biochemical measures like blood glucose and blood pressure readings) had missing values, which were imputed, potentially introducing bias.
2. Class Imbalance: Certain outcomes (e.g., hyperglycemia) suffered from extreme imbalance, leading to reduced recall despite balancing techniques.
3. Limited Feature Scope: Although the STEPS survey includes rich behavioral and physical data, it lacks genomic, dietary details, and longitudinal follow-up information, which may limit model generalizability.
4. Cross-sectional Nature: The dataset is cross-sectional, making it difficult to establish causal relationships or assess progression over time.
5. External Validity: Models were trained and validated only on Bangladeshi survey data. Their performance may differ in other LMIC contexts without external validation.

Conclusion

This study demonstrates a comprehensive machine learning framework for predicting NCD risk factors in Bangladesh using the nationally representative WHO STEPS 2018 dataset. By employing robust preprocessing, balancing strategies, hyperparameter optimization, and explainable AI, we developed interpretable models capable of predicting diabetes, hypertension, overweight/obesity, and hyperglycemia with clinically meaningful accuracy.

CatBoost and XGBoost were particularly effective for diabetes and hypertension, LightGBM and ensembles excelled for obesity, while SVM and Logistic Regression offered moderate predictive value for hyperglycemia. SHAP analysis confirmed the importance of modifiable risk factors such as BMI, blood pressure, diet, and physical activity, providing actionable insights for early intervention.

Overall, our findings highlight the potential of ML-driven health risk prediction to support public health decision-making in Bangladesh. With future improvements in data integration and validation, such frameworks can contribute significantly to personalized prevention strategies and the reduction of NCD burden in low- and middle-income countries.

References

1. S. Birdi, R. Rabet, S. Durant, A. Patel, T. Vosoughi, M. Shergill, C. Costanian, C. P. Ziegler, S. Ali, D. Buckeridge, M. Ghassemi, J. Gibson, A. John-Baptiste, J. Macklin, M. McCradden, K. McKenzie, S. Mishra, P. Naraei, A. Owusu-Bempah, L. Rosella, J. Shaw, R. Upshur, and A. D. Pinto, "Bias in machine learning applications to address non-communicable diseases at a population-level: a scoping review," BMC Public Health, 2024, [Online]. Available: https://doi.org/10.1186/s12889-024-21081-9.
2. Barbieri et al., "Deep Learning for Cardiovascular Disease Prediction in New Zealand Biomedical Database," 2022.
3. Lam et al., "Unsupervised Machine Learning with Wearable Data for T2D Risk Stratification: UK Biobank Study," 2021.
4. Ravaut et al., "Deep Learning and Survival Models for Type 2 Diabetes Prediction in Canadian Administrative Data," 2021.
5. Birk et al., "GLMM and Random Forest for T2D with Longitudinal Survey Data," 2021.
6. Alexander et al., "Natural Language Processing from Electronic Medical Records for Lung Cancer Prediction in Australia," 2019.
7. Nikola Savic, "Systematic Review: ML Applications to Population Health and Behavioral Risk Factors," (no journal specified), [Online]. Available: https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/attachments/79429923/d6fc4fbb-fed0-42fd-a766-f0e4f7235289/fmed-1-1506641.pdf.
8. "Pima Indians Diabetes Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
9. "Public Health Dataset for Heart Disease Prediction," Kaggle.
10. "Histopathological Image Dataset for Lung Cancer Classification," Kaggle.
11. "Br35H Brain Tumor Dataset," Kaggle.
12. "HARP (Heart Attack Risk Prediction) Dataset," Kaggle.
13. "Cleveland Heart Disease Dataset," [Online].
14. BaizidKhoorshid Riaz, Md Z. Islam, A. N. M. S. Islam, M. M. Zaman, Md A. Hossain, Md M. Rahman, F. Khanam, K. M. B. Amin, and I. N. Noor, "National prevalence and distribution of NCD risk factors among Bangladeshi adults," (no formal journal cited).
15. National Health and Nutrition Examination Survey, CDC.
16. Behavioral Risk Factor Surveillance System, CDC.

17. UK Biobank, [online]. Available: https://www.ukbiobank.ac.uk.
18. National Survey of Students' Health (Brazil).
19. Kaiser Permanente Dataset, [online].
20. "MIMIC-IV v2.2 database," Beth Israel Deaconess Medical Center, Boston, USA.
21. "Deep Red Fox Belief Prediction System: Deep Belief Network + Red Fox Optimization," (No formal paper cited for method; see attached summary).
22. "Framingham Heart Study Dataset," [Online].
23. "Parkinson's Disease Smartwatch Dataset," [Online].
24. "Extreme Gradient Boosting for T2D: Ajou University Hospital CDM (South Korea)," 2022.
25. "Medication Non-Adherence in Diabetic/Hypertensive Patients: ML prediction using Random Forest on CIMAS Data (Zimbabwe)," 2024.
26. "Artificial Neural Network for Diabetes (dataset summary only)," (no formal paper cited).
27. "Tight Frame Theory and Quantum Computing for Medical Image Denoising/Big Data (method summary, no paper cited)."
28. "Integration of Transformer-Based Models with MIMIC-IV EHR for Disease Prediction," (Summary, no formal citation).