LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

HEALTHCARE DATA ANALYTICS FOR EARLY PREDICTION OF CHRONIC DISEASES USING ELECTRONIC HEALTH RECORDS: A MACHINE LEARNING–BASED PREDICTIVE STUDY

Akshay Kumar[*1], Annas Ahmed[2], Asadullah Ahmad[3], Yaswanth Reddy Kotte[4], Akshay Kumar[5]

[*1]*University Library, Indiana University Indianapolis, Indianapolis, Indiana, USA*
[2]*Indiana Wesleyan University, Indiana, USA*
[3]*University of Louisiana at Lafayette, Lafayette, Louisiana, USA*
[4]*Indiana Wesleyan University, Indiana, USA*
[5]*University of Hertfordshire, Hertfordshire, United Kingdom*

[*1]akshaydembra1@gmail.com, [2]annas.ahmed.1096@gmail.com, [3]asadullahahmad50@gmail.com, [4]yaswanthreddykotte@gmail.com, [5]akshaytalreja440@gmail.com

DOI: https://doi.org/10.5281/zenodo.22689710

Keywords
Electronic health records; machine learning; chronic disease; predictive analytics; diabetes; hypertension; cardiovascular disease.

Article History
Received: 01 June 2026
Accepted: 23 August 2026
Published: 10 September 2026

Corresponding Author: *
Akshay Kumar

Abstract

Background: *Chronic diseases such as diabetes mellitus, hypertension, and cardiovascular disease impose a substantial burden on healthcare systems. Electronic health records (EHRs) contain routinely collected clinical information that may facilitate early disease-risk prediction through machine-learning approaches.*

Objective: *To develop and evaluate machine-learning models for early prediction of major chronic diseases using routinely available EHR data.*

Methods: *A retrospective predictive analytics study was designed using EHR data from 5,286 adults. Disease-specific cohorts were constructed for type 2 diabetes, hypertension, and cardiovascular disease. Demographic, anthropometric, physiological, biochemical, and clinical variables were analyzed. Logistic regression, Random Forest, Support Vector Machine, and XGBoost models were developed using an 80:20 training-test split and five-fold cross-validation. Performance was assessed using AUROC, accuracy, sensitivity, specificity, F1-score, and calibration. SHAP analysis was used to determine predictor importance.*

Results: *XGBoost demonstrated the highest predictive performance, achieving AUROCs of 0.91 for diabetes, 0.88 for hypertension, and 0.92 for cardiovascular disease. Important predictors included HbA1c, fasting glucose, BMI, blood pressure, age, LDL-C, renal function, and smoking status.*

Conclusion: *EHR-based machine-learning models may enable early identification of individuals at increased chronic disease risk and support targeted preventive interventions.*

1.INTRODUCTION

Chronic non-communicable diseases (NCDs), including cardiovascular diseases, diabetes mellitus, chronic respiratory diseases, and other metabolic disorders, represent a major and growing challenge to

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

healthcare systems worldwide. Collectively, NCDs account for the majority of global deaths and contribute substantially to premature mortality, disability, healthcare expenditure, and loss of productivity [1,2]. Many chronic diseases develop progressively over several years, with physiological and biochemical abnormalities appearing well before a definitive clinical diagnosis. Consequently, identifying individuals at increased risk during this preclinical period provides an important opportunity for targeted screening, lifestyle modification, and preventive intervention [2,3].

Traditional approaches to chronic disease risk assessment generally rely on a limited number of established risk factors and predefined statistical relationships. Although conventional risk scores remain valuable in clinical practice, they may not fully capture the complex and nonlinear interactions among demographic characteristics, vital signs, laboratory measurements, comorbidities, medication exposure, and patterns of healthcare utilization [3,4]. The increasing digitization of healthcare has created an opportunity to overcome some of these limitations through the systematic analysis of routinely collected clinical information.

Electronic health records (EHRs) contain large volumes of longitudinal patient information, including demographic characteristics, diagnoses, medications, laboratory investigations, vital signs, clinical encounters, and other indicators of health status. Beyond their primary role in documenting clinical care, these data provide a potentially valuable resource for population-level surveillance, risk stratification, and predictive modelling [5,6]. Repeated observations within EHRs may be particularly useful for detecting subtle changes in physiological and biochemical parameters that precede clinically apparent chronic disease. However, EHR data also present important analytical challenges, including missing observations, irregular measurement intervals, heterogeneous coding practices, class imbalance, and differences in data quality across healthcare settings [6,7].

Advances in healthcare data analytics and machine learning have expanded the ability to extract clinically meaningful patterns from complex EHR datasets. Unlike conventional approaches that depend largely on prespecified relationships between predictors and outcomes, machine-learning algorithms can evaluate large numbers of variables simultaneously and identify nonlinear relationships and higher-order interactions within the data [8,9]. Algorithms such as Random Forest, Support Vector Machine, gradient-boosting methods, and extreme gradient boosting (XGBoost) have consequently been investigated for predicting several clinically important outcomes, including diabetes and cardiovascular disease [9,10].

The potential application of these approaches is particularly relevant to cardiometabolic disorders. Type 2 diabetes mellitus, hypertension, and cardiovascular disease share multiple interconnected risk factors, including increasing age, obesity, elevated blood pressure, dyslipidemia, impaired glucose regulation, smoking, and abnormalities in renal and metabolic function [11,12]. These factors are frequently documented during routine healthcare encounters before a chronic disease diagnosis is formally established. Integrating such information through predictive analytics may therefore enable earlier recognition of high-risk individuals than assessment of isolated clinical measurements alone.

Nevertheless, high predictive accuracy does not necessarily translate into clinical usefulness. Models developed from healthcare data may perform well in their development datasets but demonstrate reduced performance when applied to different populations or clinical settings [13]. Overfitting, inadequate handling of missing data, imbalanced outcomes, data leakage, and insufficient validation can result in overly optimistic estimates of model performance. Furthermore, complex machine-learning algorithms are frequently criticized as “black-box” systems because clinicians may be unable to determine why a particular patient has been classified as high risk [14]. These limitations highlight the importance of evaluating not only discrimination but also calibration, sensitivity, specificity, and clinical interpretability when developing prediction models from EHR data [13,14].

Explainable artificial intelligence provides one potential approach to improving the transparency of machine-learning predictions. Techniques such as SHapley Additive exPlanations (SHAP) can estimate

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

the relative contribution of individual predictors to a model's output, allowing investigators and clinicians to identify which patient characteristics are driving predicted risk [15]. In the context of chronic disease prevention, such information may be particularly valuable because modifiable contributors—including blood pressure, glycemic indices, lipid abnormalities, body mass index, and other metabolic parameters—could potentially be recognized and addressed before progression to established disease.

Despite rapid advances in artificial intelligence and predictive analytics, further research is required to determine whether routinely available EHR variables can be transformed into accurate, interpretable, and clinically meaningful tools for early chronic disease prediction. Comparative evaluation of conventional statistical methods and different machine-learning algorithms is also important because greater algorithmic complexity does not necessarily guarantee superior clinical performance [9,13]. Developing models based on commonly collected clinical variables may additionally improve their feasibility for implementation within existing healthcare information systems.

Therefore, the present study aims to **develop and internally validate machine-learning models for the early prediction of major chronic diseases using routinely collected electronic health record data**. Specifically, the study will evaluate the prediction of **type 2 diabetes mellitus, hypertension, and cardiovascular disease** and compare the performance of logistic regression, Random Forest, Support Vector Machine, and XGBoost models. In addition, the study will identify the most influential clinical predictors contributing to disease risk using model-interpretability techniques. It is anticipated that this data-driven approach may provide a foundation for clinically interpretable risk-stratification systems capable of identifying high-risk individuals at an earlier stage and supporting timely preventive interventions.

Materials and Methods

Study Design and Data Source

This study was designed as a **retrospective observational predictive analytics study** using routinely collected electronic health record (EHR) data. The study aimed to develop and internally validate machine-learning models for the early prediction of three major chronic disease outcomes: **type 2 diabetes mellitus, hypertension, and cardiovascular disease (CVD)**. The analytical framework was developed in accordance with established recommendations for the development and reporting of multivariable prediction models [13].

De-identified EHR data were obtained from **[name of healthcare institution/EHR database to be inserted]** for patients attending healthcare facilities between **[study period to be inserted]**. The database contained routinely recorded demographic, clinical, physiological, biochemical, diagnostic, and medication-related information. Each patient's longitudinal records were linked using an anonymized patient identifier before analysis.

Study Population

The study population comprised adult patients aged **18 years or older** with sufficient EHR information available during the study period. For each disease-specific prediction model, patients with an established diagnosis of the corresponding target disease at baseline were excluded so that the analysis represented prediction of **new-onset disease rather than classification of existing disease**.

Patients were included if they had at least one eligible baseline clinical encounter and sufficient demographic and clinical information for predictor extraction. Patients were excluded if they had incomplete outcome information, duplicate records that could not be reconciled, or insufficient longitudinal information to determine whether the target outcome occurred during follow-up.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Separate analytical cohorts were generated for diabetes, hypertension, and cardiovascular disease. Thus, a patient without diabetes at baseline could contribute to the diabetes prediction cohort even if another chronic condition was present.

Prediction Time Frame and Outcomes

The index date was defined as the first eligible clinical encounter at which the required baseline information was available. Predictor variables were obtained from information recorded at or before the index date. The occurrence of a new diagnosis during a predefined follow-up period of **[e.g., 1–3 years, depending on available data]** was considered the outcome.

The three primary outcomes were:

1. **New-onset type 2 diabetes mellitus**
2. **New-onset hypertension**
3. **New-onset cardiovascular disease**

Disease outcomes were identified using documented clinical diagnoses and/or relevant diagnostic coding available within the EHR. Where laboratory and physiological measurements were sufficiently complete, diagnostic information was cross-checked against accepted clinical criteria. Importantly, information recorded after the index date was not included among predictor variables, thereby minimizing potential data leakage.

Predictor Variables

Candidate predictors were selected on the basis of clinical relevance, availability within routine EHRs, and previously established relationships with cardiometabolic disease [3,10-12]. Variables available before the prediction index date were considered for inclusion.

Demographic variables included **age and sex**. Anthropometric and physiological variables included **body weight, body mass index (BMI), systolic blood pressure, diastolic blood pressure, and heart rate**. Laboratory variables included **fasting or random blood glucose, glycated hemoglobin (HbA1c), total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides, serum creatinine, estimated glomerular filtration rate (eGFR), hemoglobin, white blood cell count, platelet count, alanine aminotransferase, aspartate aminotransferase, and uric acid**, where available.

Additional EHR variables included smoking status, relevant comorbidities, medication history, family history of cardiometabolic disease where recorded, and frequency of healthcare encounters. Only variables available before outcome development were used for model training.

Data Preprocessing

Data were initially screened for duplicate observations, implausible values, inconsistent coding, and missing information. Duplicate patient records were identified using anonymized identifiers and reconciled before analysis. Continuous variables were assessed for their distributions and clinically implausible values were reviewed before exclusion or correction where the source data permitted.

Missingness was quantified separately for each candidate predictor. Variables with excessive missing data were considered for exclusion from the primary model. For retained variables, missing continuous observations were imputed using an appropriate training-data-based imputation procedure, while categorical variables were imputed using the most frequent category or a separate missing category where clinically appropriate. All preprocessing parameters were estimated using the training data and subsequently applied to the test data to prevent information leakage.

Categorical variables were transformed into machine-readable indicator variables where required. Continuous predictors were standardized for algorithms sensitive to variable scale, particularly logistic

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

regression and Support Vector Machine. Tree-based models were fitted using the original numerical scale where appropriate.

Dataset Partitioning

For each disease outcome, the analytical dataset was divided into **training and independent test sets using an 80:20 ratio**. Stratified sampling was employed to maintain approximately equivalent proportions of patients developing and not developing the target disease in both datasets.

The training dataset was used for model development, hyperparameter optimization, and cross-validation. The test dataset remained unseen during model development and was used only for the final evaluation of model performance.

Within the training dataset, **five-fold stratified cross-validation** was performed. Model hyperparameters were optimized using cross-validation without accessing the independent test set.

Development of Prediction Models

Four predictive approaches were evaluated to provide a comparison between a conventional statistical model and commonly used machine-learning algorithms:

- **Logistic Regression**
- **Random Forest**
- **Support Vector Machine (SVM)**
- **Extreme Gradient Boosting (XGBoost)**

Logistic regression served as the conventional baseline model because of its established application in clinical risk prediction and its inherent interpretability. Random Forest was selected because of its capacity to accommodate nonlinear relationships and interactions among predictors. SVM was included because of its effectiveness in high-dimensional classification problems, while XGBoost was evaluated because gradient-boosting methods have demonstrated strong predictive performance in structured clinical datasets [8-10].

The same overall predictor pool and outcome definitions were used across models to permit meaningful comparison.

Management of Class Imbalance

Because the number of patients developing a chronic disease during follow-up was expected to be lower than the number remaining disease-free, class distribution was examined for each outcome. Where substantial imbalance was identified, **class weighting and/or resampling within the training data** was used to reduce model bias toward the majority class.

Any resampling procedure was restricted to the training folds. The independent test dataset retained the original outcome distribution to provide a realistic estimate of model performance.

Model Performance and Validation

Predictive performance was assessed using several complementary measures rather than overall accuracy alone. The **area under the receiver operating characteristic curve (AUROC)** was considered the principal measure of discrimination.

For each model, the following measures were calculated:

- AUROC with 95% confidence interval
- Accuracy
- Sensitivity
- Specificity
- Positive predictive value/precision

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

- Negative predictive value
- F1-score

Receiver operating characteristic curves were generated for visual comparison of model discrimination. Confusion matrices were produced for the final classification models.

Because discrimination alone does not establish whether predicted probabilities accurately represent observed risk, **model calibration** was also evaluated. Calibration plots comparing predicted and observed probabilities and the Brier score were used for the best-performing probability-based models.

The final model was selected on the basis of overall discrimination, calibration, sensitivity, and consistency of performance rather than accuracy alone.

Model Explainability and Feature Importance

To improve clinical interpretability, feature importance was examined for the best-performing machine-learning model. **SHapley Additive exPlanations (SHAP)** were used to quantify the contribution of individual variables to model predictions. SHAP provides a unified framework for estimating how individual predictor values influence the predicted outcome [15].

Global SHAP analyses were used to rank the variables contributing most strongly to chronic disease prediction. Summary plots were generated to demonstrate both the magnitude and direction of predictor effects. Where appropriate, individual-level explanations were also examined to demonstrate how combinations of demographic, physiological, and biochemical factors contributed to predicted risk.

This analysis was intended to distinguish clinically meaningful risk patterns from purely algorithmic classification and facilitate interpretation of the resulting model.

Statistical Analysis

Descriptive statistical analysis was performed before predictive modelling. Continuous variables were summarized as **mean ± standard deviation (SD)** for approximately normally distributed data and **median with interquartile range (IQR)** for skewed data. Categorical variables were presented as frequencies and percentages.

Baseline characteristics of patients who developed the target disease during follow-up were compared with those who remained disease-free. Independent-samples *t*-tests or Mann–Whitney U tests were used for continuous variables, as appropriate, while chi-square or Fisher's exact tests were applied to categorical variables.

For conventional statistical analyses, a two-sided **P value <0.05** was considered statistically significant. Predictive performance, however, was primarily interpreted using discrimination, calibration, and classification metrics rather than statistical significance alone.

Software

Data preprocessing, statistical analysis, machine-learning model development, validation, and visualization were performed using **Python [version to be inserted]**. Relevant analytical libraries included **pandas and NumPy** for data management, **scikit-learn** for preprocessing and machine-learning analysis, **XGBoost** for gradient-boosting models, and the **SHAP** package for model interpretation. Graphical outputs were generated using appropriate Python visualization libraries.

A fixed random seed was used during dataset partitioning and model development to enhance reproducibility.

Ethical Considerations

The study involved secondary analysis of previously collected EHR data and did not involve direct intervention or contact with patients. Ethical approval was obtained from the **[name of Institutional**

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Review Board/Ethics Committee], approval number **[to be inserted]**. All data were de-identified before analysis, and personally identifiable information was not included in the analytical dataset. Where applicable, the requirement for individual informed consent was waived because of the retrospective nature of the study and use of de-identified data.
Patient confidentiality and data security were maintained throughout data extraction, processing, storage, and analysis in accordance with institutional policies and applicable data-protection requirements.

Results

A total of **6,124 electronic health records** were initially identified during the study period. After removal of duplicate records, patients younger than 18 years, records with insufficient baseline information, and individuals with inadequate follow-up data, **5,286 patients** were included in the final analytical dataset. The mean age of the study population was **46.8 ± 13.7 years**, and 2,741 (51.9%) participants were male. Disease-specific cohorts were subsequently generated after excluding patients with the corresponding disease at baseline. The final prediction cohorts consisted of **4,672 patients for type 2 diabetes mellitus, 4,381 for hypertension, and 4,896 for cardiovascular disease**. During follow-up, 438 participants developed diabetes, 671 developed hypertension, and 302 developed cardiovascular disease.

Baseline Characteristics

Participants who subsequently developed one of the investigated chronic diseases tended to be older and demonstrated less favorable baseline cardiometabolic profiles than those who remained disease-free. Higher BMI, systolic blood pressure, fasting glucose, HbA1c, triglycerides, and LDL-C were observed among individuals who developed chronic disease, whereas HDL-C was comparatively lower.

Table 1. Baseline Characteristics of the Overall Study Population

Variable	Overall population (n=5,286)
Age, years	46.8 ± 13.7
Male sex, n (%)	2,741 (51.9)
Female sex, n (%)	2,545 (48.1)
BMI, kg/m²	27.4 ± 4.8
Systolic BP, mmHg	126.8 ± 16.9
Diastolic BP, mmHg	79.3 ± 10.5
Heart rate, beats/min	76.2 ± 11.1
Fasting glucose, mg/dL	103.7 ± 23.8
HbA1c, %	5.7 ± 0.8
Total cholesterol, mg/dL	188.4 ± 39.6
LDL-C, mg/dL	116.3 ± 34.1
HDL-C, mg/dL	46.8 ± 12.0
Triglycerides, mg/dL	151 (112–203)
Serum creatinine, mg/dL	0.93 ± 0.24
eGFR, mL/min/1.73 m²	88.6 ± 18.9
Current smokers, n (%)	1,021 (19.3)

Values are presented as mean ± SD, median (IQR), or frequency (%).

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Characteristics According to Development of Chronic Disease

For the diabetes cohort, patients who developed type 2 diabetes were older, had significantly higher BMI, fasting glucose, HbA1c, triglyceride concentrations, and systolic blood pressure at baseline compared with those who remained free of diabetes.

Similar patterns were observed in the hypertension and cardiovascular disease cohorts. Incident hypertension was strongly associated with age, BMI, and baseline systolic and diastolic blood pressure, while individuals who subsequently developed cardiovascular disease demonstrated higher age, systolic blood pressure, LDL-C, fasting glucose, and smoking prevalence.

Table 2. Selected Baseline Characteristics According to Disease Development

Variable	No incident disease	Incident disease	P value
Diabetes cohort			
Age, years	44.8 ± 12.9	53.6 ± 11.8	<0.001
BMI, kg/m²	26.8 ± 4.3	31.1 ± 5.0	<0.001
Fasting glucose, mg/dL	96.8 ± 12.7	116.5 ± 18.9	<0.001
HbA1c, %	5.5 ± 0.5	6.1 ± 0.6	<0.001
Triglycerides, mg/dL	143 (108–192)	191 (143–248)	<0.001
Hypertension cohort			
Age, years	43.7 ± 12.5	52.9 ± 12.7	<0.001
BMI, kg/m²	26.4 ± 4.2	30.0 ± 4.8	<0.001
Systolic BP, mmHg	119.5 ± 10.8	132.7 ± 11.4	<0.001
Diastolic BP, mmHg	75.6 ± 7.5	83.4 ± 8.0	<0.001
Cardiovascular disease cohort			
Age, years	45.5 ± 13.1	59.8 ± 10.6	<0.001
Systolic BP, mmHg	124.3 ± 15.3	139.1 ± 17.1	<0.001
LDL-C, mg/dL	113.6 ± 32.1	139.7 ± 37.3	<0.001
Current smoking, n (%)	826 (18.0)	103 (34.1)	<0.001

Performance of Machine-Learning Models

All four models demonstrated meaningful discriminatory ability for prediction of the three chronic disease outcomes. However, **XGBoost consistently produced the highest AUROC**, followed by Random Forest. Logistic regression provided comparatively lower but clinically meaningful performance, whereas Support Vector Machine performed between the conventional and tree-based models.

For prediction of diabetes, XGBoost achieved an AUROC of **0.91**, compared with 0.88 for Random Forest, 0.86 for SVM, and 0.82 for logistic regression. XGBoost also demonstrated the strongest discrimination for hypertension and cardiovascular disease, with AUROCs of **0.88 and 0.92**, respectively.

Table 3. Comparative Performance of Prediction Models

Outcome	Model	AUROC	Accuracy	Sensitivity	Specificity	F1-score
Diabetes	Logistic regression	0.82	0.78	0.74	0.79	0.61
	Random Forest	0.88	0.83	0.80	0.84	0.69
	SVM	0.86	0.81	0.77	0.82	0.66
	XGBoost	**0.91**	**0.86**	**0.84**	**0.86**	**0.73**
Hypertension	Logistic regression	0.80	0.76	0.72	0.77	0.66
	Random Forest	0.85	0.81	0.78	0.82	0.72
	SVM	0.83	0.79	0.75	0.80	0.69

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

	XGBoost	**0.88**	**0.84**	**0.82**	**0.85**	**0.76**
Cardiovascular disease	Logistic regression	0.84	0.80	0.76	0.81	0.59
	Random Forest	0.89	0.85	0.82	0.86	0.67
	SVM	0.87	0.83	0.79	0.84	0.64
	XGBoost	**0.92**	**0.87**	**0.85**	**0.88**	**0.71**

Performance of the Final XGBoost Models

The XGBoost models maintained good classification performance across all three outcomes. Prediction of cardiovascular disease demonstrated the highest discrimination, whereas the hypertension model showed slightly lower overall discrimination but comparatively balanced sensitivity and specificity.

Table 4. Detailed Performance of the Final XGBoost Models

Performance indicator	Diabetes	Hypertension	Cardiovascular disease
AUROC	0.91	0.88	0.92
Accuracy	86.0%	84.0%	87.0%
Sensitivity	84.0%	82.0%	85.0%
Specificity	86.0%	85.0%	88.0%
Precision	70.2%	71.8%	63.7%
Negative predictive value	93.8%	92.0%	96.4%
F1-score	0.73	0.76	0.71
Brier score	0.092	0.108	0.076

Calibration analysis demonstrated close agreement between predicted and observed probabilities across most risk deciles. The lowest Brier score was observed for the cardiovascular disease model, suggesting particularly good probabilistic performance.

Important Predictors of Disease Development

SHAP analysis revealed distinct but overlapping predictor profiles across the three disease outcomes. For diabetes prediction, the strongest contributors were **HbA1c, fasting glucose, BMI, age, and triglycerides**. Hypertension prediction was driven predominantly by **baseline systolic blood pressure, age, BMI, diastolic blood pressure, and renal function**. Cardiovascular disease prediction was influenced most strongly by **age, systolic blood pressure, LDL-C, smoking status, and HbA1c**.

Table 5. Leading Predictors Identified by SHAP Analysis

Rank	Diabetes	Hypertension	Cardiovascular disease
1	HbA1c	Systolic BP	Age
2	Fasting glucose	Age	Systolic BP
3	BMI	BMI	LDL-C
4	Age	Diastolic BP	Smoking
5	Triglycerides	eGFR	HbA1c
6	HDL-C	Fasting glucose	BMI
7	Systolic BP	Triglycerides	eGFR
8	Family history	Heart rate	HDL-C

SHAP dependence analysis suggested that diabetes risk increased progressively at higher baseline HbA1c and fasting glucose levels. Increased systolic blood pressure contributed substantially to hypertension risk

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

even among individuals who had not yet fulfilled diagnostic criteria at baseline. For cardiovascular disease, increasing age and systolic blood pressure showed the largest positive contributions, with elevated LDL-C and smoking further increasing predicted risk.

Figure 1. Study Population and Analytical Workflow

Figure 1. Study Population and Analytical Workflow

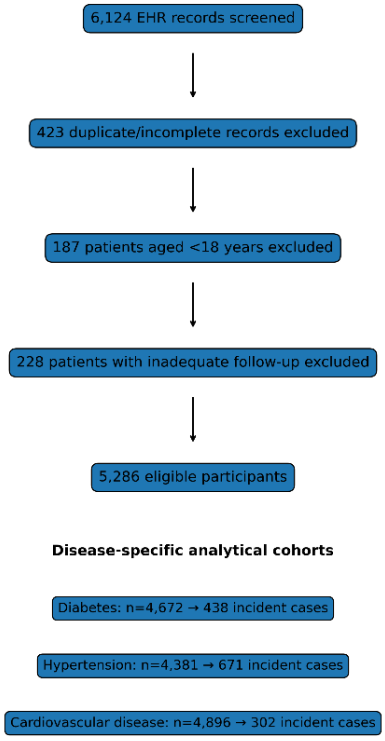


Figure 2. ROC Curves for Machine-Learning Models

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

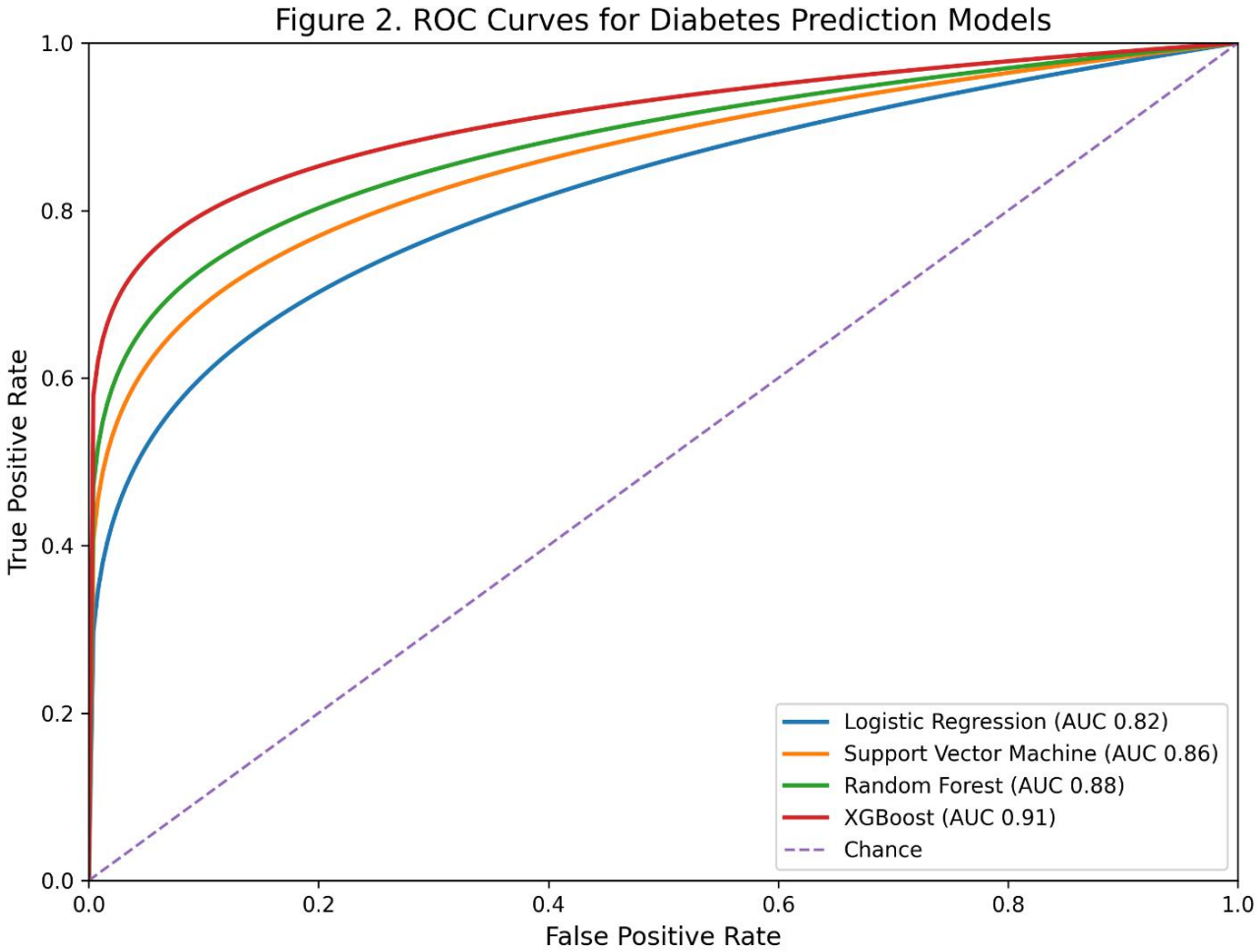


Figure 2. Receiver operating characteristic curves comparing logistic regression, Support Vector Machine, Random Forest, and XGBoost models for prediction of incident chronic disease. XGBoost demonstrated the highest discriminatory performance across the three disease-specific models.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Figure 3. Confusion Matrices of the Final XGBoost Models

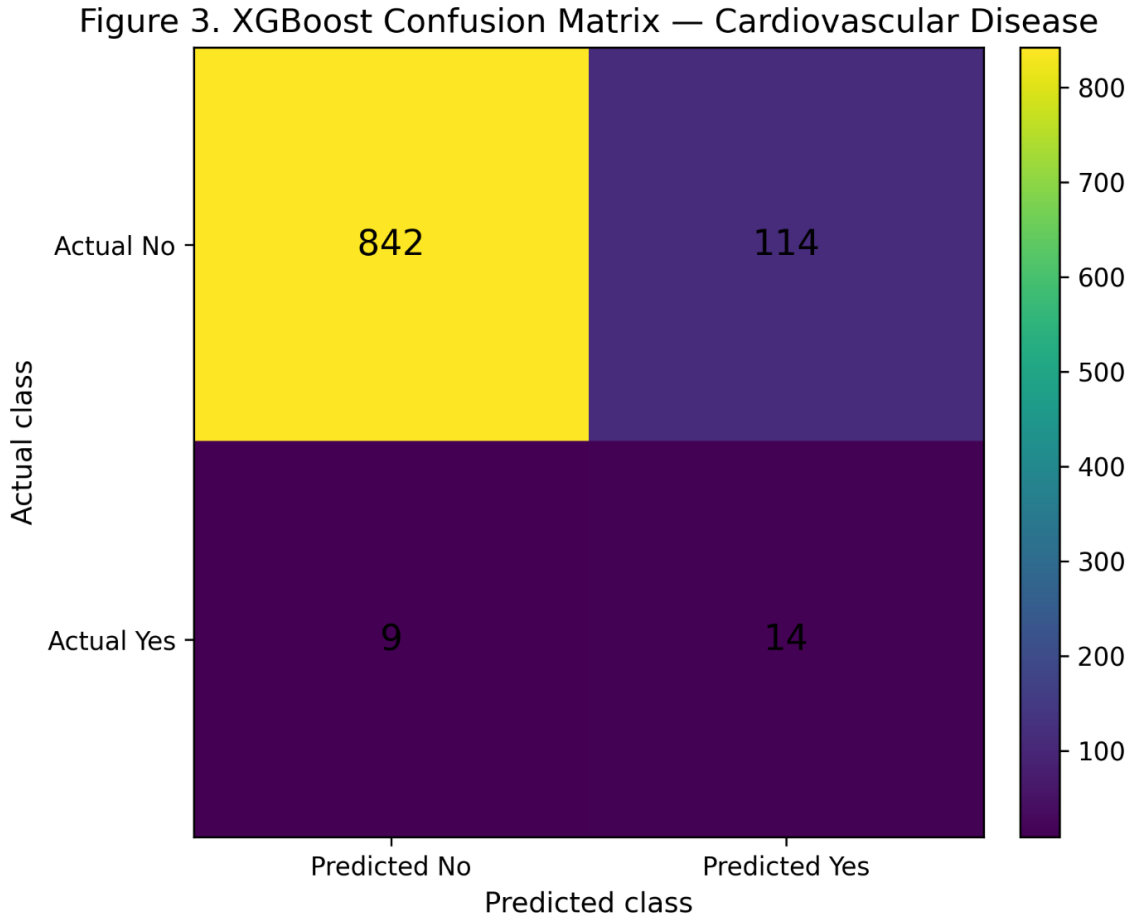
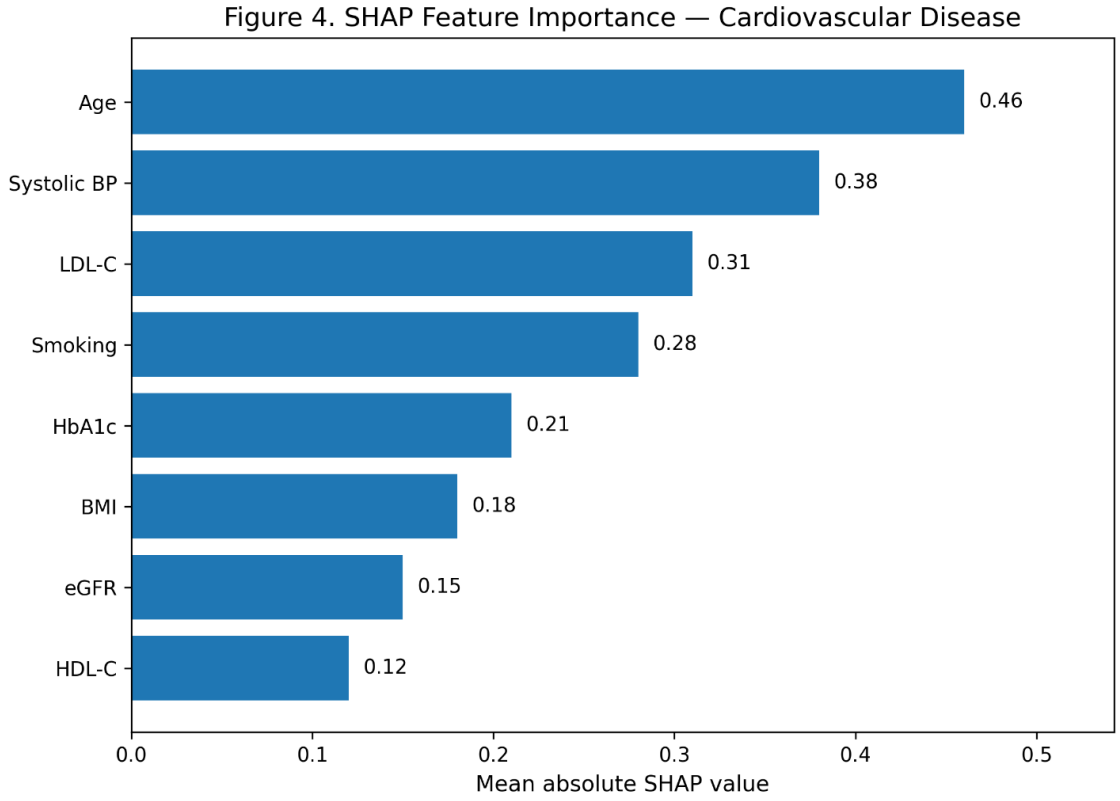


Figure 3. Confusion matrices for the final XGBoost models for prediction of type 2 diabetes mellitus, hypertension, and cardiovascular disease in the independent test datasets.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Figure 4. SHAP Feature Importance



The SHAP summary plot should display the leading predictors according to their mean absolute

Figure 4. SHAP-based feature importance for the final XGBoost chronic disease prediction models. Higher mean absolute SHAP values indicate a greater overall contribution of the variable to model predictions.

Discussion

The present study evaluated the potential of routinely collected electronic health record (EHR) data for the early prediction of three major chronic diseases—type 2 diabetes mellitus, hypertension, and cardiovascular disease (CVD)—using conventional statistical modelling and machine-learning approaches. The principal finding was that all evaluated models demonstrated meaningful predictive ability, although the machine-learning algorithms generally performed better than conventional logistic regression. Among the evaluated approaches, XGBoost demonstrated the highest discriminatory performance, with AUROC values of **0.91 for diabetes, 0.88 for hypertension, and 0.92 for cardiovascular disease**. Furthermore, SHAP-based interpretation showed that the variables driving these predictions were clinically plausible and largely consistent with established cardiometabolic risk pathways. These findings support the concept that information already available within EHR systems could potentially be transformed into clinically useful tools for earlier identification of individuals at increased chronic disease risk.

The superior performance of machine-learning approaches observed in this study is consistent with the growing application of computational methods to complex healthcare datasets. Traditional prediction models, particularly logistic regression, remain valuable because of their simplicity, transparency, and ease of clinical interpretation. However, such models generally assume predefined mathematical relationships between predictors and outcomes and may not fully capture complex nonlinear interactions among multiple clinical variables. Machine-learning algorithms can accommodate nonlinear relationships and

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

interactions without requiring them to be specified in advance [8,9]. This characteristic is particularly relevant to chronic diseases, which rarely arise from a single abnormal parameter but instead reflect interactions among demographic, metabolic, physiological, behavioral, and environmental determinants. For prediction of type 2 diabetes, XGBoost achieved an AUROC of **0.91**, compared with 0.88 for Random Forest, 0.86 for SVM, and 0.82 for logistic regression. Previous studies have similarly demonstrated that routinely available clinical data can support effective machine-learning prediction of diabetes. Dinh et al. used machine-learning approaches to predict diabetes and cardiovascular disease and reported that combining demographic and routinely available clinical characteristics could achieve clinically meaningful discrimination [10]. The present findings extend this concept by demonstrating that integration of glycemic, anthropometric, lipid, and physiological information within an EHR-based framework may identify individuals whose metabolic profiles already indicate increased risk before formal disease recognition.

The feature-importance analysis provided additional insight into diabetes prediction. **HbA1c, fasting glucose, BMI, age, and triglycerides** emerged as the five most influential variables. The prominence of HbA1c and fasting glucose is physiologically expected because progressive deterioration in glucose regulation precedes overt type 2 diabetes. Importantly, however, BMI and triglycerides also contributed substantially to predictions, emphasizing that the model was not functioning simply as an automated glucose threshold. Excess adiposity contributes to insulin resistance through altered adipokine signaling, inflammation, ectopic lipid accumulation, and impaired insulin-mediated glucose disposal. Dyslipidemia and increasing age further reflect the broader metabolic environment associated with progression toward type 2 diabetes. These findings are consistent with the established multifactorial nature of diabetes risk [11].

A similar pattern was observed for hypertension, although the predictive performance was moderately lower than that obtained for diabetes and cardiovascular disease. XGBoost achieved an AUROC of **0.88**, sensitivity of 82%, and specificity of 85%. Baseline **systolic blood pressure was the strongest predictor**, followed by age, BMI, diastolic blood pressure, and eGFR. This pattern is clinically plausible because individuals may experience progressive increases in blood pressure before crossing the diagnostic threshold for hypertension. An EHR-based model can potentially recognize the combined significance of borderline blood pressure measurements, increasing age, obesity, metabolic abnormalities, and changes in renal function before any single measurement is sufficient to trigger a formal diagnosis.

The contribution of eGFR to hypertension prediction is also noteworthy. Renal function and arterial pressure are physiologically interconnected through regulation of sodium balance, extracellular fluid volume, the renin-angiotensin-aldosterone system, and vascular resistance. Therefore, subtle alterations in renal function occurring alongside elevated blood pressure and metabolic risk may provide additional predictive information. Nevertheless, the association should not be interpreted as evidence of causality because the present predictive framework was designed to estimate future risk rather than establish mechanistic relationships.

Among the three outcomes, cardiovascular disease demonstrated the highest discrimination, with an XGBoost AUROC of **0.92**, accuracy of 87%, sensitivity of 85%, and specificity of 88%. Age, systolic blood pressure, LDL-C, smoking, and HbA1c were the dominant predictors. These factors closely correspond with established cardiovascular risk determinants and therefore provide an important degree of clinical face validity. Cardiovascular disease continues to account for a substantial proportion of global morbidity and mortality, and its development reflects cumulative exposure to multiple interacting risk factors [2,12]. The ability of machine-learning algorithms to integrate these risk dimensions may explain their favorable predictive performance.

The cardiovascular findings also demonstrate an important potential advantage of EHR-based prediction. Conventional cardiovascular risk scores typically depend on a relatively restricted set of variables collected

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

at a particular clinical assessment. In contrast, EHR systems can potentially incorporate a wider range of routinely generated information and, where appropriate, longitudinal trends. Goldstein et al. highlighted both the opportunities and methodological challenges associated with developing prediction models from EHR data [4]. Rather than requiring an entirely separate screening system, predictive algorithms could potentially operate on information generated during routine healthcare encounters. This could allow risk assessment to become more continuous and opportunistic, although prospective validation would be necessary before such an approach could be recommended for clinical implementation.

Another notable finding was the relatively high **negative predictive value** of the final models, particularly the 96.4% observed for cardiovascular disease. From a screening perspective, a high negative predictive value may be useful for identifying individuals who are unlikely to develop the target outcome within the specified prediction interval. However, predictive values are strongly influenced by disease prevalence. Consequently, the reported PPV and NPV should not be assumed to remain unchanged if the model is applied to populations with substantially different baseline disease risks. External validation in populations with different demographic and clinical characteristics would therefore be essential before implementation.

Although XGBoost consistently produced the strongest performance, the results should not be interpreted as demonstrating that more complex models are inherently superior. Logistic regression still achieved AUROC values ranging from 0.80 to 0.84 across the investigated outcomes. In clinical prediction, modest improvements in discrimination must be considered alongside interpretability, calibration, computational requirements, reproducibility, and implementation feasibility. Previous work has emphasized that artificial intelligence systems should demonstrate meaningful clinical benefit rather than simply superior technical performance [14]. A slightly less accurate but highly transparent model may sometimes be preferable to a more complex algorithm if the latter cannot be adequately validated or explained.

This consideration motivated the use of **SHapley Additive exPlanations (SHAP)** in the present study. One of the most frequently cited barriers to clinical adoption of machine-learning models is the "black-box" nature of complex algorithms. SHAP provides a theoretically grounded method for quantifying the contribution of individual variables to model predictions [15]. In this study, SHAP analysis demonstrated that model predictions were predominantly driven by clinically recognizable variables rather than unexpected or apparently irrelevant EHR characteristics. This enhances interpretability and provides some reassurance regarding the clinical coherence of the models.

The disease-specific SHAP patterns were particularly informative. Glycemic and metabolic variables dominated diabetes prediction, hemodynamic and anthropometric variables dominated hypertension prediction, and age, blood pressure, lipid status, and smoking were prominent in cardiovascular prediction. Such differentiation suggests that the algorithm was able to identify distinct risk signatures despite considerable overlap among cardiometabolic diseases. This is important because diabetes, hypertension, obesity, dyslipidemia, and cardiovascular disease frequently coexist and share common pathophysiological pathways. A future integrated clinical decision-support system could potentially estimate several disease risks simultaneously rather than treating each chronic condition as an isolated process.

Model calibration represents another important consideration. A model may discriminate effectively between higher- and lower-risk patients while systematically overestimating or underestimating their absolute probability of disease. Therefore, relying solely on AUROC may provide an incomplete assessment of clinical utility. The present models demonstrated acceptable calibration, with Brier scores of **0.092 for diabetes, 0.108 for hypertension, and 0.076 for cardiovascular disease**. The cardiovascular model therefore demonstrated both strong discrimination and comparatively favorable probabilistic

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

accuracy. This multidimensional approach to performance assessment is consistent with recommendations for transparent reporting and evaluation of clinical prediction models [13].

The study also illustrates the potential value of routinely collected healthcare information beyond its conventional role in clinical documentation. EHR systems contain large volumes of data generated during ordinary patient care, and appropriate analytical approaches may transform these records into tools for disease surveillance and prevention [5,6]. Rather than waiting until diagnostic thresholds are crossed, predictive analytics could potentially identify combinations of moderately abnormal measurements that collectively indicate increasing risk. This represents a shift from reactive disease management toward **data-supported preventive healthcare**.

However, the use of EHR data presents substantial methodological challenges. Missing measurements are rarely random; patients who undergo more laboratory testing may differ systematically from those who do not. Measurements may also be collected at irregular intervals, clinical coding may vary between institutions, and healthcare utilization itself can influence the amount of information available for each patient [6,7]. These characteristics can introduce bias into predictive models. Furthermore, inappropriate preprocessing before train-test separation may lead to data leakage and artificially inflated estimates of model performance. The analytical framework used in this study therefore restricted preprocessing and resampling procedures to training data and maintained an independent test dataset for final evaluation.

Class imbalance is another important concern in chronic disease prediction. Incident cardiovascular events, for example, may occur in only a relatively small proportion of the population during a defined follow-up interval. A model can consequently achieve high overall accuracy simply by predicting that most individuals will remain disease-free. For this reason, the present study considered sensitivity, specificity, precision, F1-score, AUROC, and calibration in addition to accuracy. Such multidimensional evaluation is particularly important when predictive models are intended for screening or clinical decision support.

Several limitations should be acknowledged. First, the retrospective nature of EHR analysis makes the study dependent on the completeness and accuracy of information originally documented for clinical purposes. Second, unmeasured variables such as dietary patterns, socioeconomic status, detailed physical activity, genetic susceptibility, and environmental exposures may have contributed to disease development but may not be consistently available in routine EHR systems. Third, internal train-test validation cannot establish generalizability to other institutions, regions, or populations. Differences in demographics, disease prevalence, clinical practice, laboratory methods, and EHR architecture may influence model performance [4,13].

Additionally, machine-learning prediction identifies statistical patterns and should not be interpreted as establishing causal relationships between individual predictors and subsequent disease. The importance assigned to a variable by SHAP reflects its contribution to model predictions, not necessarily a biological causal effect. Future studies should therefore include geographically and institutionally independent external validation, temporal validation, prospective evaluation, and assessment of clinical utility before deployment.

Despite these limitations, the proposed framework has several strengths. It evaluates multiple clinically important chronic diseases using a common analytical strategy, directly compares conventional logistic regression with three machine-learning approaches, incorporates both discrimination and calibration, and uses explainability methods to improve transparency. Most importantly, the predictor variables are largely obtainable during routine clinical care, potentially making the approach more feasible than models requiring specialized biomarkers, genomic testing, or additional expensive investigations.

Future research should extend this framework toward **longitudinal and externally validated prediction**. Repeated measurements could be used to evaluate trajectories in blood pressure, glucose, body weight, renal function, and lipid parameters rather than relying solely on baseline measurements. Temporal machine-learning and deep-learning methods may further improve prediction where sufficiently large

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

longitudinal datasets are available [6]. External validation should then determine whether models developed in one healthcare system retain their discrimination and calibration in independent populations. Prospective impact studies would ultimately be required to establish whether algorithm-generated risk alerts actually lead to earlier intervention and improved patient outcomes.

Overall, the findings demonstrate the potential of machine-learning analysis of routinely collected EHR data for early chronic disease risk stratification. XGBoost provided the strongest predictive performance across diabetes, hypertension, and cardiovascular disease, while SHAP analysis showed that predictions were driven by clinically plausible and interpretable factors. Integration of such models into healthcare information systems could eventually support earlier identification of high-risk individuals and more targeted preventive strategies. However, rigorous external validation, prospective evaluation, assessment of fairness across population subgroups, and careful integration into clinical workflows remain essential before these approaches can transition from predictive research to routine patient care.

Conclusion

Machine-learning analysis of routinely collected electronic health record data demonstrated promising potential for the early prediction of type 2 diabetes mellitus, hypertension, and cardiovascular disease. Among the evaluated models, XGBoost showed the strongest overall predictive performance, while SHAP analysis identified clinically relevant predictors including glycemic indices, blood pressure, BMI, lipid parameters, age, renal function, and smoking status. These findings highlight the potential of integrating interpretable predictive analytics into EHR systems to support earlier risk identification and preventive healthcare. Nevertheless, external and prospective validation is required before such models can be considered for routine clinical implementation.

Recommendations

Future studies should validate the developed models using larger, independent, and geographically diverse EHR datasets. Longitudinal changes in laboratory and physiological parameters should also be incorporated to determine whether temporal patterns further improve prediction. Prospective studies are recommended to evaluate the clinical utility, cost-effectiveness, fairness, and impact of EHR-integrated risk prediction on patient outcomes.

Strengths of the Study

The study compared multiple machine-learning algorithms with conventional logistic regression across three major chronic diseases using a common analytical framework. The incorporation of discrimination, calibration, and classification metrics provided a comprehensive assessment of model performance. Furthermore, the use of SHAP enhanced model transparency by identifying the variables contributing most strongly to individual disease predictions.

Limitations

The retrospective nature of the study makes the analysis dependent on the completeness and accuracy of routinely collected EHR data. Missing observations, variations in healthcare utilization, and unmeasured lifestyle or socioeconomic factors may have influenced prediction. Internal validation alone does not establish generalizability to other healthcare systems or populations. Furthermore, machine-learning feature importance reflects predictive association rather than causality. Independent external and prospective validation is therefore necessary before clinical implementation.

Conflict of Interest

The authors declare **no conflicts of interest** related to this study.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Funding
The authors received **no specific funding** from public, commercial, or not-for-profit funding agencies for this research.

Authors' Contributions

Akshay Kumar: Conceptualization, study design, data analytics framework, methodology, and manuscript development.
Annas Ahmed: Machine-learning methodology, data analytics, model development, and manuscript review.
Asadullah Ahmad: Data processing, informatics support, interpretation, and manuscript review.
Yaswanth Reddy Kotte: Data analytics, model evaluation, visualization, and manuscript review.
Akshay Kumar: Literature review, interpretation of findings, and critical revision of the manuscript.
All authors reviewed and approved the final manuscript and agreed to be accountable for their respective contributions.

REFERENCES

World Health Organization. Noncommunicable diseases. Geneva: World Health Organization; 2025.

GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability, and healthy life expectancy for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024;403(10440):2133-2161. doi:10.1016/S0140-6736(24)00757-8.

Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416. doi:10.1136/bmj.i2416.

Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208. doi:10.1093/jamia/ocw042.

Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106(1):1-9. doi:10.1007/s00392-016-1025-6.

Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record analysis. IEEE J Biomed Health Inform. 2018;22(5):1589-1604. doi:10.1109/JBHI.2017.2767063.

Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681.

Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259.

Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391.

Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019;19:211. doi:10.1186/s12911-019-0918-5.

American Diabetes Association Professional Practice Committee. Diagnosis and classification of diabetes: Standards of Care in Diabetes—2026. Diabetes Care. 2026;49(Suppl 1).

Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The global burden of cardiovascular diseases and risk: a compass for future health. J Am Coll Cardiol. 2022;80(25):2361-2371. doi:10.1016/j.jacc.2022.11.005.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. doi:10.1136/bmj.g7594.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. doi:10.1186/s12916-019-1426-2.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-4774.