Machine Learning Regressions for Non-Communicable Diseases: A Review of Diagnostic Applications and Clinical Intervention Strategies

Panida Aroonsirichoke[1], Subij Shakya[2], Manaporn Chatchamni[3], Ravinan Thatsiriniratkul[3], Jittrapat Kongsup[4] and Pichit Boonkrong[*1]

[1] College of Biomedical Engineering, Rangsit University, Pathum Thani 10400, Thailand
[2] Department of Food and Nutrition, University of Helsinki, Helsinki 00014, Finland
[3] School of Nursing, Rangsit University, Pathum Thani 10400, Thailand
[4] Department of Surgery, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[*]Corresponding author, E-mail: pichit.bk@rsu.ac.th

Abstract

Non-communicable diseases (NCDs) represent a critical global health crisis, necessitating proactive risk prediction through continuous assessment. This review systematically examines 15 studies utilizing machine learning (ML) regression to forecast NCD variables from Electronic Health Records (EHRs) and clinical data. Moving beyond binary classification, these studies analyze continuous outcomes such as time to renal replacement therapy (days), HbA1c levels, blood glucose (mg/dL), systolic blood pressure (mmHg), body weight (kg), medical cost (USD), etc. The reviewed research encompasses diverse sample sizes, ranging from small clinical cohorts of 12 patients to large-scale populations exceeding 63,000 participants. While advanced MLs consistently demonstrate superior accuracy in capturing non-linear relationships, traditional models like multiple linear and logistic regression remain competitive for simpler clinical predictors. Key pre-processing steps, including data splitting, K-fold cross-validation, imputation, and normalization, are identified as essential for model stability. Model performance is rigorously assessed using standard metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), to gauge predictive accuracy and explain variance. Techniques such as SHAP (SHapley Additive exPlanations) values and partial dependence plots are highlighted as essential tools for visualizing how individual clinical factors influence specific risk scores. The integration of these frameworks addresses the critical trade-off between the high accuracy of complex algorithms and the transparency required for clinical trust. This paper outlines a roadmap for integrating transparent, high-precision ML frameworks into clinical decision support systems to enable personalized, data-driven interventions.

Keywords: *clinical data, electronic health records, factor analysis, healthcare analytics, regression models*

1. Introduction

Non-communicable diseases (NCDs) present a devastating and growing global health crisis, responsible for millions of premature deaths annually (Agboyibor et al., 2026; Jarintanan et al., 2024; World Health Organization, 2026). Accurate and proactive risk prediction is essential for implementing timely preventive strategies. The application of machine learning (ML) regression has provided a pivotal shift in NCD prediction, moving analyses beyond simple binary classification toward a continuous, quantitative assessment of risk or prognosis (Adlung et al., 2021; Fujihara et al., 2023; Simmachan et al., 2025; Zheng & Yu, 2021).

ML regression techniques have emerged as transformative tools, offering the ability to process complex physiological datasets and predict disease progression with high accuracy (Boonkrong et al., 2025b; Yang et al., 2025). By shifting from traditional statistical methods to sophisticated algorithmic frameworks, clinicians can better quantify risk factors, enhancing personalized treatment strategies and improving long-term patient outcomes (Boonkrong, et al., 2025b; Kristensen et al., 2023; Mehta et al., 2025b; Simmachan et al., 2025). This capability is paramount, as chronic diseases such as cardiovascular disease, diabetes, and kidney disease manifest across a spectrum of physiological indicators rather than at a single discrete threshold. Contemporary studies leverage vast, multi-modal healthcare datasets, including longitudinal Electronic Health Records (EHRs), laboratory values, and patient-reported outcomes, to train these models. The literature is increasingly dominated by comparisons between traditional linear models and advanced non-

linear regression algorithms. While some researchers, such as Nusinovici et al. (2020), argue that "*Binary Logistic Regression (BLR) was as good as ML for predicting major chronic diseases*" when using simple clinical predictors, others demonstrate the necessity of more advanced architectures. For example, Geng et al. (2024) proposed a regionally generalized ML framework for census-enabled, multifactor NCD analyses, utilizing Cubist regression to estimate disease prevalence with high precision. The integration of biochemical markers into regression models has further improved the granularity of chronic disease staging. Liu et al. (2021) conducted a comparative study and development of advanced ML tools to predict nonalcoholic fatty liver disease, demonstrating how models like XGBoost can effectively map clinical factors to a continuous Fatty Liver Index (FLI). Complex ensemble techniques, including Decision Trees (DTs), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting (GB), consistently demonstrate superior predictive accuracy, minimizing metrics such as Root Mean Squared Error (RMSE) when forecasting outcomes such as future Body Mass Index (BMI) or risk indices. However, the gain in predictive performance often comes at the cost of model transparency. Significant methodological advancements have been made in applying ML regression to NCD prognosis; however, substantial challenges persist across the data lifecycle. As highlighted by Haq et al. (2020), Adlung et al. (2021), Soriano-Valdez et al. (2021), Yadav et al. (2023) refining hyper-parameters and feature selection remains critical for managing the inherent complexities of clinical datasets.

The future of chronic care hinges on translating complex, non-transparent algorithms into robust, integrated clinical decision support systems. This shift will enable personalized, proactive risk forecasting, empowering clinicians to deploy timely interventions. Currently, the lack of a structured synthesis of ML regression in NCDs hinders standardized clinical translation; this review addresses that gap by harmonizing diverse methodologies into a unified framework for continuous outcome prediction. By bridging this divide, researchers can ensure that high-precision models remain both robust and explainable for safe clinical adoption.

2. Objectives

Standardizing NCD risk forecasting requires bridging the gap between data inconsistencies and opaque algorithms to ensure robust, interpretable clinical adoption. Thus, the objectives of this paper are listed as follows:

1) To provide a guideline on how preprocessing techniques like normalization and scaling significantly influence the reliability of regression outcomes

2) To identify the most significant clinical and lifestyle features, guiding researchers on which variables to prioritize in future NCD dataset

3) To provide a roadmap for using interpretability tools to ensure that model predictions are grounded in biological reality before patient application

4) To evaluate the integration of ensemble architectures into clinical decision support systems for capturing complex, non-linear disease progression patterns

3. Methodology

The conceptual framework for this systematic review follows a rigorous pipeline designed to bridge computational ML accuracy with clinical utility, especially NCDs. To ensure the selected literature aligned with the review's four core objectives, specific eligibility criteria were established. The systematic methodology is structured into the following 6 stages:

1) Keyword Identification and Searching: The process begins by identifying primary keywords, e.g., clinical data, chronic diseases, risk factors, machine learning, regression models, etc. Researchers then search multi-modal healthcare databases, including longitudinal Electronic Health Records (EHRs) and clinical trial registries to identify relevant papers on NCDs like cardiovascular disease, hypertension, diabetes, obesity, and kidney disease.

2) Screening and Selection: Studies are screened based on their focus on continuous NCD risk scores rather than binary classification. Table 1 defines the scope for selecting 15 studies based on timeframe, data

Proceedings of RSU International Research Conference (RSUCON-2026)

sources, ML regression methods, and continuous outcomes. This review specifically examines 15 studies that utilize ML regression to forecast variables from clinical and demographic data.

3) *Data Extraction and Pre-processing*: Relevant information is extracted regarding data cleaning, normalization, and handling missing values. Methodological steps, e.g., data splitting, *K*-fold cross-validation, feature engineering and feature scaling are documented to assess model stability.

4) *Analytical Comparison*: The methodology involves evaluating the performance of traditional models (e.g., Multiple Linear Regression) against advanced non-linear algorithms like DTs, KNN, XGB, RF, SVR, etc.

5) *Quality Assessment and Interpretability*: Models are reviewed for their use of interpretability tools like SHAP, LIME, ICE values or partial dependence plots to ensure predictions are grounded in biological reality.

6) *Synthesis and Summarization*: Finally, findings are synthesized into Table 2, detailing input features, ML models, evaluation metrics (RMSE, MAE, R^2), and clinical outcomes for each study.

Table 1 Inclusion and exclusion criteria for the systematic review of ML regression in NCDs.

Inclusion Criteria	Exclusion Criteria
1. Studies published specifically between the years 2020 and 2025.	1. Studies that focus solely on binary classification (e.g., disease presence vs. absence)
2. Use of EHRs, clinical cohorts, or population-based surveys.	2. Research that does not use ML or statistical regression as the primary method.
3. Reporting of standard metrics, e.g., RMSE, MAE, or R^2 for model evaluation.	3. Qualitative or descriptive studies that do not provide quantitative metrics.
4. Studies utilizing ML regression for predicting continuous NCD variables like HbA1c, blood pressure, BMI, or medical costs.	4. Data sources that are strictly non-clinical or unrelated to NCD risk forecasting.

4. Conceptual Framework in ML Regression for NCDs Prediction

This conceptual framework delineates the systematic pipeline for predicting continuous NCD outcomes. It integrates diverse data sources with rigorous pre-processing and feature selection to optimize ML regression performance as shown in Figure 1. By prioritizing model interpretability and robust evaluation, the framework bridges computational accuracy with actionable clinical implications for personalized healthcare.

4.1 Data Description

The datasets utilized for predicting NCDs typically comprise multi-modal health records, including continuous metrics, demographic information, and lifestyle factors. These datasets often originate from long-term EHRs, clinical trials, or population-based health surveys like the large cohort used for spirometry references (Haq et al., 2020; Mehta et al., 2025b). Specific studies leverage census-enabled data and environmental factors to estimate disease prevalence across regions. Considering the feature set, the ML inputs can be nominal, ordinal, count or continuous (Adlung et al., 2021; Haq et al., 2020; Soriano-Valdez et al., 2021). Key continuous outcomes of interest often include blood pressure (mmHg), glucose levels (mg/dL), body temperature (°C), SpO_2 (%), BMI (kg/m^2), heart rate (bpm), blood urea nitrogen (BUN, mg/dL), HbA1c (%), Albumin (g/dL) etc. Reliable NCD risk assessment is essential, requiring diverse datasets from clinical records and longitudinal EHRs to ensure model stability.

4.2 Pre-processing

To enhance the reliability of NCD regression models, rigorous data pre-processing is essential to mitigate noise, sparsity, and systemic biases inherent in clinical datasets. Initial stages prioritize data integrity through robust cleaning, outlier detection, and the imputation of missing values to address incomplete records (Haliduola et al., 2022; Nusinovici et al., 2020; Sreejith et al., 2020). Feature engineering, including standardization and scaling, was applied to maintain numerical stability and minimize the impact of extreme

variance on systolic blood pressure predictions. Furthermore, signal quality is refined via smoothing and interpolation, while dimensionality is optimized through Principal Component Analysis (PCA) and iterative feature selection (Gárate-Escamila et al., 2020; Geng et al., 2024; Zhang et al., 2021). To address class distribution disparities, the Synthetic Minority Over-sampling Technique (SMOTE) is frequently deployed (Kristensen et al., 2023; Zaizar-Fregoso et al., 2023). Combined with train/test splitting and *K*-fold cross-validation, this comprehensive workflow ensures that the resulting predictive models are both statistically sound and generalizable across diverse populations.

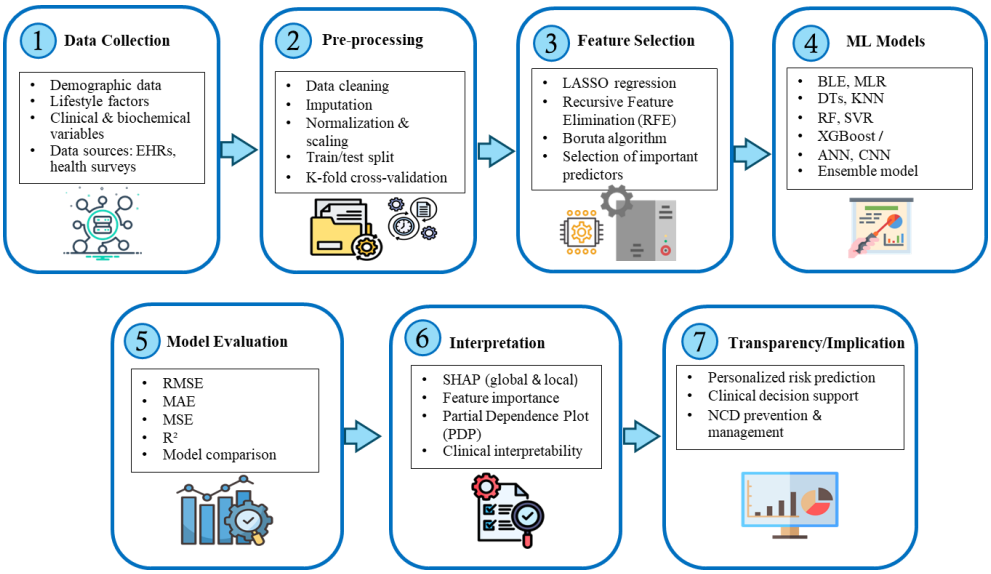

Figure 1 Methodological framework in ML regression

4.3 Feature Selection

Feature selection serves as a cornerstone in predictive modeling for NCDs, where the high dimensionality of clinical data often necessitates the isolation of parsimonious variable sets (Haq et al., 2020; Sreejith et al., 2020). By utilizing sophisticated wrappers like Recursive Feature Elimination (RFE) and the Boruta algorithm, researchers can systematically identify the most potent clinical, demographic, and lifestyle predictors while discarding redundant noise. Furthermore, regularization frameworks, e.g., LASSO (L1) and Ridge (L2) regression, are instrumental in managing multicollinearity; these methods apply penalty terms to less significant coefficients, effectively shrinking them to enhance model interpretability and prevent overfitting (Gárate-Escamila et al., 2020; Simmachan et al., 2025). This rigorous filtering ensures that only critical biomarkers, e.g., sodium intake for hypertension or specific biochemical markers for hepatic pathology, inform the final regression analysis.

4.4 ML Regression

To map the intricate relationships between multifaceted input features and continuous health outcomes, a diverse array of machine learning architectures is deployed within the NCD domain. Traditional approaches, such as Binary Logistic Regression (BLR), Multiple Linear Regression (MLR), and K-Nearest Neighbors (KNN), provide foundational benchmarks for predicting metrics like regional prevalence (Huang et al., 2022; Mehta et al., 2021a; Nusinovici et al., 2020; Taloba et al., 2022). These are often augmented by regularization techniques such as LASSO and Ridge regression to enhance sparsity and stability. For capturing complex, non-linear dependencies, researchers frequently utilize SVR, DTs, and Gaussian Process Regression (GPR). Advanced ensemble methods, including Random Forest (RF), Stochastic Gradient Boosting (SGB), and Classification and Regression Trees (CART), are increasingly favored for their robust

predictive power (Boonkrong et al., 2025b; Das et al., 2024; Huang et al., 2022; Kristensen et al., 2023; Simmachan et al., 2025). High-performance frameworks like XGBoost, LightGBM, and Stacking Regressors further refine these outcomes by iteratively minimizing loss functions. For high-dimensional or temporal clinical data, specialized architectures such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks are employed to track disease progression over time (Nusinovici et al., 2020; Zheng & Yu, 2021). Finally, niche methodologies like Cubist regression and Heterogeneous Mixture Learning Technology (HMLT) offer high-precision estimations for specific physiological markers, e.g., BMI and biochemical fluctuations (Fujihara et al., 2023). While no single "*gold standard*" exists, a comparison of outcomes like HbA1c and blood glucose reveals that ensemble methods (XGBoost, RF) and temporal models (LSTM) consistently outperform traditional regression models in predictive accuracy.

4.5 Evaluation

Evaluation of these models involves a systematic assessment using standard metrics to gauge predictive accuracy and variance explanation. The most common metrics reported are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 values (Boonkrong & Simmachan, 2026; Simmachan et al., 2025). Some specialized studies also utilize the Brier Score or AUC when assessing risk probabilities, or use adjusted R^2 to account for model complexity. Robustness is further verified through rigorous spatial cross-validation or train-validation splits to ensure that the findings are generalizable across different healthcare settings and populations.

Table 2 ML Regression Evaluation Metrics for NCDs

Metric	Mathematical Expression	Clinical Meaning and Use Case
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	MAE represents the average magnitude of error in the original clinical units, e.g., glucose (mg/dL) or blood pressure (mmHg). It provides clinicians with a straightforward understanding of the typical prediction deviation for daily monitoring of weight changes (kg) in obesity management or medical cost estimations in USD.
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert \hat{y}_i\rvert} \times 100$	MAPE expresses prediction error as a percentage, allowing for easy comparison of model accuracy across different NCD variables regardless of their scale. It is particularly useful to evaluate the relative accuracy of glycemic-related outcomes or percentage changes in ejection fraction for heart failure patients.
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Squaring the differences between actual and predicted health metrics, MSE serves as a mathematical foundation for optimizing complex non-linear algorithms. Commonly, it is used as an internal benchmark to refine models like Ridge regression or XGBoost when mapping biochemical factors to liver disease indices.
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	RMSE heavily penalizes large clinical outliers, making it essential for ensuring patient safety where extreme prediction errors could lead to dangerous interventions. It is a critical guideline for validating models predicting time to renal replacement therapy or HbA1c levels, where high-precision forecasting is paramount.
RAE	$\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{\sum_{i=1}^{n}\lvert \hat{y}_i - \bar{y}\rvert}$	RAE provides a dimensionless ratio comparing the model's error to a simple mean-based baseline, indicating how much better the ML technique performs than a naive guess. It is a valuable information tool for researchers assessing the predictive power of clinical data against complex urine albumin-creatinine ratios in diabetes.
R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	R^2 quantifies the proportion of clinical variance explained by the model, indicating how well demographic and lifestyle features account for disease progression. It serves as a gold standard for assessing the clinical utility of models predicting systolic blood pressure or disease prevalence across regional populations.

Table 3: Summary of 15 recent studies using ML regression models for predicating NCD variables

No.	Author (Year)	Disease	Samples	Input Features (*X*)	Output Variable (*Y*)	Pre-processing	ML Model	Evaluation Metrics	Interpretation
1	Nusinovici et al. (2020) [34]	CKD, CVD, Diabetes	6,762	Clinical predictors (Age, BP, BMI)	Risk Scores/Probability	Imputation, Centering, Scaling	BLR, RF, XGBoost, KNN, CNN	AUC, Brier Score	Feature Ranking
2	Liu et al. (2021)	NAFLD	25	24 Clinical/Biochemical factors	Fatty Liver Index	Feature ranking	XGBoost, Ridge	AUC, MSE	Feature Ranking
3	Mehta et al. (2021a)	CVD	299	Echocardiogram metrics, Heart rate	Ejection fraction (%)	Standardization	MLR, KNN	R^2	SHAP
4	Zheng & Yu (2021)	Hypertension/ CVD	250	Clinical, Lifestyle	Systolic BP (mmHg)	Iterative feature selection	ANN, LR, SVM, DTR, GPR	BHS grade, AAMI	N/A
5	Zhang et al. (2021)	Diabetes	12	Diabetes data, Vital signs, Physical activities	Blood glucose (mg/dL)	Normalization, Smoothing	LSTM,	MAE, RMSE, MAPE, R^2	Clarke error grid
6	Huang et al. (2022)	Diabetes	1,147	Demographics, Blood chem, Vital sign	Urine albumin-creatinine ratio	N/A	MLR, RF, SGB, CART, XGBoost	MAPE, SMAPE, RAE	Feature Ranking
7	Taloba et al. (2022)	Obesity	24,353	Demographics, Lifestyles	Medical cost (USD)	N/A	MLR	RMSE, MAE, MSE, MAPE, R^2	Clinical significance
8	Chen et al. (2023)	Hypertension	400	Age, BMI, Sodium intake, Genetics	Systolic BP (mmHg)	SHAP	LightGBM, LASSO	R^2, MSE	SHAP
9	Fujihara et al. (2023)	Obesity	12,021	Demographics, BMI, Lifestyle, Annual health data	Body weight change over 3 years (kg.)	Normalization	HMLT	RSME	HMLT
10	Kristensen et al. (2023)	COPD	23,433	Demographics, Health data	FEV1, FVC (continuous lung function)	Imputation, Outlier removal, SMOTE	MLR, RF	R^2	Feature importance
11	Zaizar-Fregoso et al. (2023)	Diabetes	63,776	Demographic, Clinical/ Biochemical variables	Glycemic-related outcome	Feature selection, Normalization, SMOTE	XGBoost	RMSE, MAE, R^2	OR, SHAP
12	Das et al. (2024)	Diabetes	520	Insulin, Glucose, Skin thickness	HbA1c Levels (%)	Scaling	RF, SVR	RMSE, MAE	N/A
13	Okita et al. (2024)	CKD	135	EMR, Baseline eGFR, Laboratory data	Time to renal replacement therapy (Days)	Encoding, Standardization	LASSO	R^2	LASSO, SHAP
14	Geng et al. (2024)	Asthma, Diabetes, etc.	311	Census data, Socioeconomic, Environmental	Disease prevalence (%)	Smoothing, Interpolation, Scaling, PCA	Cubist	R^2, RMSE	Box-Cox transformation
15	Simmachan et al. (2025)	Psoriasis	149	Demographics, PASI, Psychological stress	DLQI score (Continuous)	Penalized feature selection	RF, SVR	R^2, RSME, MAPE	SHAP

4.6 Interpreting Feature Importance

Effective model interpretation is paramount for bridging the gap between high-performance computational outputs and clinical utility. Beyond standard feature ranking and feature importance metrics, which provide a macro-level view of variable contributions, advanced techniques like SHAP (SHapley Additive exPlanations) are utilized to decompose the specific influence of biomarkers, such as BMI or genetic predispositions, on individual hypertension risk scores (Boonkrong et al., 2025b; Rengasamy et al., 2021; Saarela & Jauhiainen, 2021; Simmachan et al., 2025; Yang et al., 2025). To ensure that ML decisions align with biological reality, researchers employ Odds Ratios (OR) and LASSO coefficients to quantify risk magnitude (Nusinovici et al., 2020; Okita et al., 2024; Zaizar-Fregoso et al., 2023). Furthermore, specialized evaluation frameworks such as the Clarke Error Grid assess the clinical safety of predictions, while the Box-Cox transformation and Heterogeneous Mixture Learning Technology (HMLT) ensure that data distributions remain interpretable (Fujihara et al., 2023; Geng et al., 2024). By centering clinical significance, these methodologies provide a transparent, evidence-based rationale that empowers medical professionals to integrate machine learning insights safely into personalized patient care.

4.7 Clinical Implication

The integration of ML regression models into clinical practice offers significant potential for personalized medicine and proactive risk management. By accurately forecasting health metrics such as HbA1c levels or time to renal replacement therapy, these tools enable clinicians to deploy targeted interventions earlier (Adlung et al., 2021; Feng et al., 2024; Soriano-Valdez et al., 2021; Sreejith et al., 2020). This shift from binary classification to continuous assessment allows for more granular disease staging and optimized resource allocation. Crucially, the transparency provided by SHAP and feature ranking ensures that computational results are directly interpretable by clinicians. These methodologies allow medical professionals to validate model logic against established biological realities, ensuring that predictions for conditions like hypertension or diabetes are grounded in clinical significance (Boonkrong et al., 2025b; Chen et al., 2023; Geng et al., 2024; Zheng & Yu, 2021).

5. Data and Methodological Challenges

Implementing machine learning for clinical regression involves navigating a complex landscape of data integrity and model validation. The following sections explore the critical balance required to transform raw EHR and census data into reliable predictive tools. Ensuring these algorithms offer actionable medical insights requires a multifaceted approach: managing vast feature sets, demystifying complex models, and applying rigorous imputation and bias-mitigation techniques to guarantee both reliability and transparency.

5.1 Data Quality and Availability

Predicting NCDs relies on diverse medical data types, including structured clinical records, unstructured physician notes, medical images (MRI/CT), and physiological signals (ECG/EEG), all of which are frequently plagued by high dimensionality and missing values (Boonkrong et al., 2025a; Mavaie et al., 2023; Simmachan et al., 2025; Spooner et al., 2020; Zheng & Yu, 2021). The reliance on diverse sources such as Longitudinal Electronic Health Records (EHRs) and census data often leads to issues of data sparsity and entry errors. Regarding data quality and availability issues, medical datasets are often fragmented, combining structured EHRs with unstructured clinical notes and high-dimensional genomic data, leading to the "*curse of dimensionality*" and overfitting. Missing data remains a critical barrier, stemming from equipment failure or inconsistent patient follow-ups and necessitating complex imputation that may introduce bias. Furthermore, human error, e.g., manual entry mistakes or subjective clinical coding, degrades the ground truth. These challenges are compounded by data silos and strict privacy regulations that limit the large-scale availability of diverse datasets. Consequently, models frequently struggle with class imbalance, where rare disease events are overshadowed by healthy samples, compromising predictive reliability. As demonstrated by Yadav et al. (2023) and Geng et al. (2024), robust pre-processing and imputation are essential to mitigate these inconsistencies and ensure model reliability. A significant barrier to the real-world deployment of these

Proceedings of RSU International Research Conference (RSUCON-2026)

ML models lacks of technical interoperability between fragmented healthcare systems. To move from research to clinical decision support (CDSS), future frameworks must adopt standardized protocols like HL7 FHIR (Fast Healthcare Interoperability Resources) to ensure seamless data exchange between diverse Electronic Health Records (EHRs) and AI-driven diagnostic tools.

5.2 Dimensional Feature Selection

To address high dimensionality in NCD prediction, researchers employ feature selection techniques to isolate the most discriminative variables from complex clinical records, medical images, physiological signals, and genomic data (Mavaie et al., 2023; Nusinovici et al., 2020; Spooner et al., 2020; Wyss et al., 2022). Filter methods use statistical tests like Chi-square (χ^2), Information Gain, and Pearson Correlation to rank features independently of a model, effectively discarding noise. Wrapper methods, such as Recursive Feature Elimination (RFE), Genetic Algorithms (GA), and Boruta, which uses "*shadow features*" to identify all truly relevant variables, evaluate feature subsets to capture complex dependencies. Embedded methods integrate selection into training. For tree-based models, this includes Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI) to quantify feature contributions (Boonkrong, et al., 2025a; Simmachan et al., 2025). Additionally, LASSO and Elastic Net penalize coefficients to refine the feature set. These techniques reduce time complexity and mitigate the non-linear effect of irrelevant features, improving model interpretability and predictive accuracy. Managing extensive clinical datasets requires sophisticated techniques like LASSO or Ridge regression to isolate influential predictors, such as systolic blood pressure or biochemical markers. While these methods help manage complexity, as seen in the work of Chen et al. (2023), they must be carefully applied to avoid overlooking non-linear feature interactions.

5.3 Generalization vs. Overfitting

In NCD prediction, achieving generalization while avoiding overfitting is critical for model reliability. Overfitting occurs when a model memorizes noise, outliers, or human errors in high-dimensional datasets, e.g., genomic markers or physiological signals, instead of identifying underlying clinical patterns. While such models perform perfectly on small training cohorts, they fail when applied to unseen populations. To ensure clinical utility, researchers employ cross-validation, regularization (LASSO/Ridge), and Dimensional Feature Selection (Boruta, MDA/MDI) to capture universal biological trends (Boonkrong et al., 2025a; Simmachan et al., 2025). Rigorous validation using metrics like Adjusted R^2 is essential to distinguish genuine patterns from dataset-specific noise. As highlighted by Mehta et al. (2021a) and Geng et al. (2024), maintaining consistent performance across diverse populations through regionally generalized frameworks is vital to prevent biased risk estimates and ensure stability across demographic settings. Beyond initial validation, continuous monitoring for data drift is vital, as ML models trained on 2020 datasets may suffer from performance decay by 2026 due to evolving patient lifestyles or updated laboratory equipment standards.

5.4 The Interpretability-Accuracy Trade-off

In predicting NCDs, a persistent challenge exists in the trade-off between the high predictive accuracy of complex models and their inherent lack of transparency. High-performing models, like ANN, CNN and XGBoost, achieve superior accuracy by capturing intricate, non-linear relationships in signals and images but often remain opaque. To bridge this gap, researchers utilize global methods like Partial Dependence Plots (PDP) and Permutation Feature Importance (PFI) to rank variable impacts across datasets. To extract feature importance from top-tier models, advanced techniques such as SHAP and LIME explain individual patient risks, while Grad-CAM provides visual heatmaps for diagnostic imaging (Rengasamy et al., 2021; Saarela & Jauhiainen, 2021; Simmachan et al., 2025). Additionally, tree-based ensembles utilize MDI and MDA to quantify feature contributions. Integrating these explainable AI (XAI) tools allows clinicians to validate complex predictions against medical knowledge, ensuring high accuracy without sacrificing the transparency required for safe clinical adoption. To resolve this, researchers must prioritize

Proceedings of RSU International Research Conference (RSUCON-2026)

the integration of interpretability frameworks to ensure that high-accuracy models remain both robust and explainable for medical professionals.

5.5 Ethical Concerns and Bias Mitigation

Implementing smart healthcare systems for NCD prediction requires navigating complex ethical concerns and legal frameworks. Under regulations like Thailand's PDPA or the GDPR, sensitive health data must be handled with strict informed consent, purpose limitation, and anonymization to ensure patient autonomy (Gerke et al., 2020; McGraw & Mandl, 2021). Bias mitigation is equally critical; models trained on skewed datasets can perpetuate healthcare disparities, necessitating techniques like reweighting and fairness-aware machine learning to ensure equitable outcomes across diverse demographics. Robust policies must mandate routine bias audits and transparency in algorithmic decision-making. By aligning technical innovation with law and ethical principles, systems can foster public trust while providing stable, fair, and legally compliant risk assessments. To further enhance privacy and data availability, future frameworks should adopt federated learning, allowing models to be trained across decentralized hospital servers without transferring sensitive patient records (Kairouz & McMahan, 2021; Rieke et al., 2020; Wyss et al., 2022). Additionally, the generation of synthetic data can mitigate class imbalances and data sparsity, providing robust, anonymized datasets for training high-precision NCD regressors while strictly adhering to PDPA and GDPR mandates (Gerke et al., 2020; Yadav et al., 2023).

6. Conclusion

The integration of ML regression models marks a transformative shift in the predictive management of NCDs, moving beyond binary classification to a more nuanced, continuous assessment of health risks. By leveraging high-dimensional datasets from EHRs and census data, these models effectively forecast critical health metrics such as disease prevalence, progression scores, and biochemical markers. The review highlights that while traditional models like BLR and MLR remain strong baselines for simple clinical predictors, advanced ML regressors, including DTs, KNN, SVR, RF and XGB, offer superior predictive accuracy for complex, non-linear relationships. Practical clinical utility depends on successfully navigating technical barriers, e.g., data sparsity and overfitting, while simultaneously demystifying the opaque nature of high-performance ensemble models. Future research must prioritize model interpretability and external validation to bridge the gap between computational performance and clinical utility.

7. Acknowledgements

The authors extend their sincere gratitude to the anonymous reviewers for their insightful feedback and constructive critiques, which significantly enhanced the quality of this work. We also gratefully acknowledge the financial support provided by the Research Institute of Rangsit University for our participation in the RSU International Research Conference 2026.

8. References

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. *Med*, *2*(6), 642-665. https://doi.org/10.1016/j.medj.2021.04.006

Agboyibor, K. M., Nambiema, A., Golestani, A., Okeibunor, J., Diallo, C. B. B., Jouven, X., ... & Empana, J. P. (2026). Prevalence, time trends and associated factors of adult overweight and obesity in 36 countries in the WHO African region from 2003 to 2022: a study of 54 WHO STEPS surveys representing 156 million adults. *BMJ Global Health*, *11*(1), Article e019988. https://doi.org/10.1136/bmjgh-2025-019988

Boonkrong, P., & Simmachan, T. (2026). A comparative analytical framework for modeling road fatalities with count regression techniques. *Transportation Research Interdisciplinary Perspectives*, *36*, Article 101807. https://doi.org/10.1016/j.trip.2025.101807

Boonkrong, P., Shakya, S., Kraunamkam, W., & Simmachan, T. (2025a). Dietary patterns and psoriasis severity in Thai patients: a machine learning approach for small sample data. *Scientific Reports*, *15*(1), Article 33088. https://doi.org/10.1038/s41598-025-17657-z

Boonkrong, P., Shakya, S., Yang, J., & Simmachan, T. (2025b). Risk factors in males and females for disease classification based on International Classification of Diseases, 10th Revision codes. *Engineering Proceedings*, *108*(1), Article 26. https://doi.org/10.3390/engproc2025108026

Chen, Y., Fan, S., Qiu, Y., & Deng, Z. (2023). SHAP-Guided Feature Selection Enhances Stacked Gradient-Boosting Regression for Radar-Based Cuffless Blood Pressure Estimation. *SSRN*, 5967439.

Das, D. K., Chowdhury, S., & Hossain, M. M. (2024, October). A comparative analysis on a diabetic disease prediction considering various machine learning algorithms. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, Kaski, Nepal. https://doi.org/10.1109/ICSES63445.2024.10763383

Feng, Z., Chen, Y. A., Guo, Y., & Lyu, J. (2024). Deciphering the environmental chemical basis of muscle quality decline by interpretable machine learning models. *The American Journal of Clinical Nutrition*, *120*(2), 407-418. https://doi.org/10.1016/j.ajcnut.2024.05.022

Fujihara, K., Yamada Harada, M., Horikawa, C., Iwanaga, M., Tanaka, H., Nomura, H., ... & Sone, H. (2023). Machine learning approach to predict body weight in adults. *Frontiers in public health*, *11*, Article 1090146. https://doi.org/10.3389/fpubh.2023.1090146

Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, Article 100330. https://doi.org/10.1016/j.imu.2020.100330

Geng, K., Thota, S., & Kataria, A. (2024). A Regionally Generalized Machine Learning Framework Towards Census-Enabled Multi-Factor Non-Communicable Disease Analyses. *2024 3rd International Conference on Artificial Intelligence and Software Engineering (ICAISE)* (pp. 57-65). IEEE, Singapore. https://doi.org/10.1109/ICAISE65384.2024.00017

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial intelligence in healthcare*, 295-336). https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Haliduola, H. N., Bretz, F., & Mansmann, U. (2022). Missing data imputation using utility-based regression and sampling approaches. *Computer Methods and Programs in Biomedicine*, *226*, Article 107172. https://doi.org/10.1016/j.cmpb.2022.107172

Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., ... & Ali, A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. *Sensors*, *20*(9), Article 2649. https://doi.org/10.3390/s20092649

Huang, L. Y., Chen, F. Y., Jhou, M. J., Kuo, C. H., Wu, C. Z., Lu, C. H., ... & Lu, C. J. (2022). Comparing multiple linear regression and machine learning in predicting diabetic urine albumin–creatinine ratio in a 4-year follow-up study. *Journal of Clinical Medicine*, *11*(13), Article 3661. https://doi.org/10.3390/jcm11133661

Jarintanan, P., Singh, N., Suthienkul, O., & Boonkrong, P. (2024). Established and emerging risk factors of stroke in asian countries: A systematic review. *Thai Journal of Public Health*, *54*(2), 952-967.

Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning*, *14*(1-2), 1-210. https://doi.org/10.1561/9781680837896

Kristensen, K., Olesen, P. H., Roerbaek, A. K., Nielsen, L., Hansen, H. K., Cichosz, S. L., ... & Hejlesen, O. (2023). Using random forest machine learning on data from a large, representative cohort of the general population improves clinical spirometry references. *The Clinical Respiratory Journal*, *17*(8), 819-828. https://doi.org/10.1111/crj.13662

Liu, Y. X., Liu, X., Cen, C., Li, X., Liu, J. M., Ming, Z. Y., ... & Zheng, S. S. (2021). Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. *Hepatobiliary & Pancreatic Diseases International*, *20*(5), 409-415. https://doi.org/10.1016/j.hbpd.2021.08.004

Mavaie, P., Holder, L., & Skinner, M. K. (2023). Hybrid deep learning approach to improve classification of low-volume high-dimensional data. *BMC bioinformatics*, *24*(1), Article 419.

McGraw, D., & Mandl, K. D. (2021). Privacy protections to encourage use of health-relevant digital data in a learning health system. *NPJ digital medicine*, *4*(1), Article 2. https://doi.org/10.1038/s41746-020-00362-8

Mehta, D., Naik, A., Kaul, R., Mehta, P., & Bide, P. J. (2021a). Death by heart failure prediction using ML algorithms. *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-5). IEEE, NaviMumbai, India. https://doi.org/10.1109/ICNTE51185.2021.9487652

Mehta, S., Huey, S. L., Fahim, S. M., Sinha, S., Rajagopalan, K., Ahmed, T., ... & Finkelstein, J. L. (2025b). Advances in artificial intelligence and precision nutrition approaches to improve maternal and child health in low resource settings. *Nature communications*, *16*(1), Article 7673. https://doi.org/10.1038/s41467-025-62985-3

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56-69.

Okita, J., Nakata, T., Uchida, H., Kudo, A., Fukuda, A., Ueno, T., ... & Shibata, H. (2024a). Development and validation of a machine learning model to predict time to renal replacement therapy in patients with chronic kidney disease. *BMC nephrology*, *25*(1), Article 101. https://doi.org/10.1186/s12882-024-03527-9

Rengasamy, D., Rothwell, B. C., & Figueredo, G. P. (2021). Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Applied Sciences*, *11*(24), Article 11854. https://doi.org/10.3390/app112411854

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), Article 119. https://doi.org/10.1038/s41746-020-00323-1

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2), Article 272. https://doi.org/10.1007/s42452-021-04148-9

Simmachan, T., Lerdpraserdpakorn, N., Deesrisu, J., Sriwipat, C., Shakya, S., & Boonkrong, P. (2025). A penalized regression and machine learning approach for quality-of-life prediction in psoriasis patients. *Healthcare Analytics*, *8*, Article 100417. https://doi.org/10.1016/j.health.2025.100417

Soriano-Valdez, D., Pelaez-Ballestas, I., Manrique de Lara, A., & Gastelum-Strozzi, A. (2021). The basics of data, big data, and machine learning in clinical practice. *Clinical Rheumatology*, *40*(1), 11-23. https://doi.org/10.1007/s10067-020-05196-z

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, *10*(1), Article 20410. https://doi.org/10.1038/s41598-020-77220-w

Sreejith, S., Nehemiah, H. K., & Kannan, A. (2020). Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. *Computers in Biology and Medicine*, *126*, Article 103991. https://doi.org/10.1016/j.compbiomed.2020.103991

Taloba, A. I., Abd El-Aziz, R. M., Alshanbari, H. M., & El-Bagoury, A. A. H. (2022). Estimation and prediction of hospitalization and medical care costs using regression in machine learning. *Journal of Healthcare Engineering*, *2022*(1), Article 7969220. https://doi.org/10.1155/2022/7969220

World Health Organization. (2026). *From Strategy to Actions: Prioritizing NCD prevention and control Regional workshop to advance NCD prevention and control in the WHO South-East Asia Region, 28–30 October 2025 Jaipur, India* (No. SEA-NCD-117). World Health Organization. Regional Office for South-East Asia.

Wyss, R., Yanover, C., El-Hay, T., Bennett, D., Platt, R. W., Zullo, A. R., ... & Lin, K. J. (2022). Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database

studies: An overview of the current literature. *Pharmacoepidemiology and drug safety*, *31*(9), 932-943. https://doi.org/10.1002/pds.5500

Yadav, P., Sharma, S. C., Mahadeva, R., & Patole, S. P. (2023). Exploring hyper-parameters and feature selection for predicting non-communicable chronic disease using stacking classifier. *IEEE Access*, *11*, 80030-80055. https://doi.org/10.1109/ACCESS.2023.3299332

Yang, J., Simmachan, T., Shakya, S., & Boonkrong, P. (2025). Classification of infectious and parasitic diseases by smart healthcare system. *Engineering Proceedings*, *108*(1), Article 14.

Zaizar-Fregoso, S. A., Lara-Esqueda, A., Hernández-Suarez, C. M., Delgado-Enciso, J., Garcia-Nevares, A., Canseco-Avila, L. M., ... & Delgado-Enciso, I. (2023). Using Artificial Intelligence to Develop a Multivariate Model with a Machine Learning Model to Predict Complications in Mexican Diabetic Patients without Arterial Hypertension (National Nested Case-Control Study): Metformin and Elevated Normal Blood Pressure Are Risk Factors, and Obesity Is Protective. *Journal of Diabetes Research*, *2023*(1), Article 8898958. https://doi.org/10.1155/2023/8898958

Zhang, M., Flores, K. B., & Tran, H. T. (2021). Deep learning and regression approaches to forecasting blood glucose levels for type 1 diabetes. *Biomedical Signal Processing and Control*, *69*, Article 102923. https://doi.org/10.1016/j.bspc.2021.102923

Zheng, J., & Yu, Z. (2021). A novel machine learning-based systolic blood pressure predicting model. *Journal of Nanomaterials*, *2021*(1), Article 9934998.