

Article

Explainable Machine Learning for Vitamin D-Related Chronic Disease Risk Prediction Using MIMIC-III

Salma Chebbawi ^{1,2}, Mohamed Tabaa ^{1,*}  and Hassan Badir ² ¹ Pluridisciplinary Laboratory of Research and Innovation (LPRI), Ecole Marocaine des Sciences de l'Ingénieur, Casablanca 20190, Morocco; chebbawi.salma@etu.uae.ac.ma² IDS Team, ENSA, University Abdel Malek Essaadi, Tanger 90000, Morocco; hassan.badir@uae.ac.ma

* Correspondence: m.tabaa@emsi.ma

Abstract

The prevalence of vitamin D deficiency has been noted to be of great concern to the population due to its association with multiple chronic ailments, such as those that affect the cardiovascular system, diabetes, bone diseases, and the immune system, and even some cancerous forms. The underlying mechanics are not well understood, but epidemiological and clinical research demonstrates notable correlations, correlated to different degrees by population, based on genetic, environmental, and lifestyle differences. The use of advanced machine learning methods in this study aimed to examine the association between serum vitamin D levels and chronic disease risk using the de-identified electronic health records from Intensive Care Unit (ICU) patients in the MIMIC-III database. A final cohort of 38,712 unique patients was selected after applying predefined inclusion and exclusion criteria, the data being used for machine learning model development and evaluation. To predict the status of chronic diseases based on vitamin D levels, supervised machine learning models were used: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). Data preprocessing involved strategies such as missing data imputation, normalization, and feature engineering to improve model training. The accuracy, precision, recall, F1-score, and receiver operating characteristic curve (ROC-AUC) were used to assess model performance. LR performed the best with the ROC-AUC of 92.5% and accuracy of 85.2%, and SVM was close behind the ROC-AUC: 90.9 percent. The chronic disease was associated with 2.1 times higher risk of chronic disease among people with severe vitamin D deficiency ($p < 0.001$), which was in agreement with earlier observational data. RF did not work as well (ROC-AUC: 87.7%) because it required higher dimensionality data, e.g., genetic and lifestyle parameters. The study shows the possible use of AI-based diagnostics in preventive medicine. As an example, presented vitamin D testing of high-risk population (e.g., increasing tests in old age, obesity, or dark-skinned population) adapted to ML-based risk assessment as well as supplementation plans could enhance this prevention.



Academic Editor: Douglas O'Shaughnessy

Received: 18 May 2026

Revised: 10 June 2026

Accepted: 13 June 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: vitamin D deficiency; chronic disease prediction; machine learning; MIMIC-III; precision medicine; healthcare AI

1. Introduction

The main cause of mortality in the world is non-communicable diseases, which include cardiovascular disease, diabetes, osteoporosis, and cancer. The World Health Organization

report reveals that these conditions kill about 74 percent of the world population [1], pointing to the dire need to develop better early detection and prevention programs. One of the biological risk factors that can be modified to develop chronic diseases is the lack of vitamin D, which has gained more significance in terms of immune regulation, metabolic processes, and inflammatory processes. Clinically, vitamin D deficiency is typically assumed as levels of 25-hydroxyvitamin D [25(OH)D] in serum lower than 20 ng/mL, a state that is approximated to impact over one billion people worldwide. Several epidemiological studies have indicated that a deficiency of vitamin D is associated with a high risk of cardiometabolic diseases, autoimmune diseases, and inflammatory diseases.

Although this is becoming increasingly clear-cut, the vitamin D status and chronic disease outcomes, when it comes to vitamin D, are a poorly understood and even a controversial relationship. These factors that have been confounding in observational studies are typically very significant, including obesity, systemic inflammation, nutrient intake, or lifestyle factors, such as physical activity and sun exposure. Further, the relationship between vitamin D and disease risk may be non-linear and threshold-based and is difficult to model using conventional statistical models. These limitations underscore the need for complex analytic models that can model the interaction of complex biomarkers and disease on large datasets of clinical data.

The problems can be addressed by recent advances in machine learning (ML), which facilitate data-driven modeling of high-dimensional clinical data and identification of non-linear relationships between biological measures and disease outcomes. In the electronic health record (EHR) analysis, risk assessment, and clinical forecasting activity, ML techniques have proven to be extremely effective. However, there are limited studies testing the predictive power of vitamin D levels via multiple machine learning models by using large, critical care datasets, while simultaneously evaluating the interpretability and clinical utility of the models.

Therefore, a crucial knowledge gap exists regarding whether vitamin D can be used as a reliable predictor of risk for chronic disease through the use of advanced machine learning methods on large-scale clinical data. Also, little information exists comparing the performance of various machine learning algorithms in this scenario and their application in clinical decision-support systems. Filling this gap is critical to converting epidemiological associations into tools that can be operationalized to aid early risk stratification, tailored preventive interventions, and better healthcare outcomes. To this end, the present study examines the predictive power of serum vitamin D levels on the risk of chronic disease in a large critical care database (MIMIC-III) using a range of machine learning algorithms. Thus, the predictive capacity of serum vitamin D concentration for the risk of chronic disease was evaluated in the present study using multiple machine learning methods within a large critical care database (MIMIC-III).

To explore the predictive value of serum 25-hydroxyvitamin D [25(OH)D] level for chronic disease risk, the largest publicly available electronic health record (EHR) repository, MIMIC-III (MESONET MICU Support), is used. This study addresses a supervised classification problem with serum vitamin D levels used as input features, along with patient demographic and clinical information obtained from patient records. The outcome variable is chronic disease status, which is a binary variable based on patients' diagnostic information. For modeling this relationship, four supervised machine learning algorithms were implemented and compared, namely Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). The models were chosen to represent a variety of statistical, kernel-based, ensemble, and distance-based learning models.

This work has threefold objectives: Firstly, to estimate the predictive value of vitamin D as an independent biomarker in the risk model of chronic diseases. Secondly, to compare the output of several machine learning algorithms with the aim of determining models that have clinical translation potential. Lastly, to assess the interpretability of the models and the biological plausibility of the models using feature importance.

Through the combination of machine learning methods with massive datasets of EHR, this research proposes to investigate the possibility of scalable, data-driven risk stratification methods, according to vitamin D status. The approaches can help to better identify the risk of chronic disease, as well as build AI-enhanced preventive care models.

2. Background

2.1. Vitamin D Metabolism and Biological Functions

Vitamin D, a fat-soluble vitamin and secosteroid, is central to human physiology, coordinating a complex diversity of biological functions via its two-pronged route of endogenous production and oral intake. The synthesis of vitamin D begins in the epidermis, where the ultraviolet B (UVB) radiation (290,315 nm) acts as a catalyst in upgrading the 7-dehydrocholesterol into a derivative known as previtamin D₃, which further isomerizes with no supplementary radiation to form cholecalciferol (vitamin D₃). Ergocalciferol (vitamin D₂) can be obtained through the diet from plant sources and foods like fatty fish, egg yolks, and fortified milk and dairy products, whilst vitamin D₃ is received from foods of animal origin. After absorption, both forms are carried through chylomicrons to the liver, where they are 25-hydroxylated by CYP450 enzymes (CYP2R1 enzyme predominant) to form 25-hydroxyvitamin D [25(OH)D], the major blood-borne metabolite and the clinical indicator of vitamin D status. The precursor 25(OH)D is then further hydroxylated in the kidneys to form 1,25-dihydroxyvitamin D [1,25(OH)₂D], the biologically active form active in the kidneys mediated by an enzyme known as 1 α -hydroxylase (Cyp27B1). Its effects could be mediated by binding with the vitamin D receptor (VDR), a nuclear transcription factor that is ubiquitously expressed in different tissues. When activated, the VDR constitutes a heterodimer with the retinoid X receptor (RXR), regulating the expression of more than 900 genes on a spectrum of varied physiological pathways.

In addition to its classical role in calcium-phosphate metabolism in an attempt to sustain skeletal health, vitamin D has a pleiotropic effect that significantly impacts whole-body health. Beyond its long-established role in calcium and phosphate absorption in the small intestine and bone mineralization, recent research emphasizes its immunomodulatory, anti-inflammatory, and metabolic potential. The biologically active vitamin D metabolite 1,25(OH)₂D is known to balance innate and adaptive immunity by down-regulating the production of pro-inflammatory cytokines (e.g., tumor necrosis factor- α [TNF- α]). The twofold activity enhances host defense against infections and reduces excessive inflammatory responses, which are associated with autoimmune diseases. Vitamin D boosts insulin response and sensitivity in metabolic tissues by increasing insulin secretion and sensitivity via activation of VDR in pancreatic beta cells and skeletal muscle, which modulates calcium-dependent processes that are essential in glucose uptake and metabolism. Moreover, 1,25(OH)₂D modulates the process of cellular differentiation and proliferation, which prevents the proliferation of abnormal cells in the breast, colon, and prostate, suggesting the possibility of vitamin D being involved in cancer prevention.

The clinical consequences of a combined vitamin D deficiency, with a high serum concentration of serum 25-hydroxyvitamin D of less than 20 ng/mL, are strong, as epidemiological evidence has been found that associates deficiency with a range of chronic pathologies, like heart diseases, type 2 diabetes, diseases of the immune system, and some cancers. These associations occur at a molecular level and are complex, including VDR

polymorphisms, epigenetic changes, and interaction with environmental factors, including UVB exposure and dietary intake. Indicatively, a lack of vitamin D alters the molecular renin–angiotensin–aldosterone system (RAAS), facilitating endothelial dysfunctions and stiffening of arteries, increasing the risk of cardiovascular disease. In the same manner, the failure in the VDR signaling in immune cells can worsen inflammatory cascades, inducing diseases such as rheumatoid arthritis and multiple sclerosis.

The complexity of the interaction of these mechanisms explains why sophisticated analytical models are required to determine the non-linear interactions between vitamin D status and disease pathogenesis, as is done in this paper. Given its combined role in the biological web, vitamin D is a risk factor that can be addressed, and its deficiency is an eventual sign of predisposition to chronic diseases in all world populations.

2.2. Epidemiological Links to Chronic Diseases

Comprehensive epidemiological studies have supported the fact that low serum 25-hydroxyvitamin D [25(OH)D] concentrations of less than 20 ng/mL are consequential in contributing to a severe intensification in the possibility of a variety of chronic diseases, which can include cardiovascular disease (CVD), type 2 diabetes mellitus (T2DM), autoimmune disease, as well as certain malignancies. This is a widespread inadequacy, which hits more than 1 billion people in the world and has significant implications in the pathogenesis of disease via complex molecular and physical pathways, making vitamin D a critical modifiable risk factor. Low levels of 25(OH)D in the CVD setting trigger the onset of endothelial dysfunction, increased arterial stiffness, and development of hypertension by derailing the renin–angiotensin–aldosterone system (RAAS). This impairment adds to the disproportion of vascular homeostasis, resulting in stimulation of atherogenesis and an increased risk of cardiovascular disease. Meta-analyses demonstrate that people with serum 25(OH)D levels in the lower range (less than 15 ng/mL) incur a 50 percent greater risk of developing a myocardial infarction than those who retain adequate amounts (greater than 30 ng/mL); other studies point towards a 30,350 percent increased odds of heart failure and stroke in subpar populations.

Comprehensive epidemiological studies have supported the fact that low concentrations of serum 25-hydroxyvitamin D [25(OH)D], less than 20 ng/mL, have a pronounced proliferative effect on the vulnerability to a cluster of chronic diseases, including cardiovascular disease (CVD), type 2 diabetes mellitus (T2DM), autoimmune conditions, and some cancers. Such a shortage, which is experienced by more than 1 billion people worldwide, causes an immense impact on the disease pathogenesis due to the complex pathways the lack of vitamin D has on the molecular and physiological levels, making it one of the most important modifiable risk factors. In CVD, low 25(OH)D is a precipitating factor to endothelial dysfunction, increased arterial stiffness, and hypertension because it affects RAAS.

This disturbance induces a vascular homeostasis imbalance that will trigger atherogenesis, leading to increased cardiovascular risk. Meta-analyses have shown that people with a serum 25(OH)D concentration below 15 ng/mL have a 50 percent higher rate of myocardial infarction than those who managed to maintain levels exceeding 30 ng/mL, and more research is showing that the deficient population has a 30 to 50 percent higher risk of heart failure and stroke.

In metabolic health, vitamin D deficiency affects the functionality of insulin and impairs insulin sensitivity of the pancreatic cells; these are important factors that determine glucose status, and hence enhance the incidence of T2DM by an estimated 30%. The observational data also highlight that people with 25(OH)D concentrations less than 20 ng/mL have slowed skeletal muscle glucose uptake and lower insulin release, and the former

occurs due to poor vitamin D receptor (VDR) activation in pancreatic islets. Such metabolic abnormalities are elaborated with systemic inflammation, which further intensifies the insulin resistance. As far as autoimmune disorders are concerned, including multiple sclerosis (MS) and rheumatoid arthritis (RA), the level of vitamin D and the condition of UVB are manageable risk factors. Persistent vitamin D alleviates the generation of autoantibodies and develops regulatory T-cell (Treg) activity that prevents inflammatory damage, which lowers the severity of the disease. Epidemiological observations show that a population that has greater UVB exposure and a level of 25(OH)D > 30 ng/mL has a reduced incidence of MS and RA by 20–40 percent, indicating the immunomodulatory possibilities of vitamin D.

In cancer, growing evidence exists indicating a protective role of vitamin D against cancer, with vitamin D modulating cancer protection by binding to VDR to inhibit the proliferation of tumor cells and angiogenesis. The insufficiency of 25(OH)D causes a 25–35 per cent worse vulnerability to breast, colorectal, and prostate malignancy, and mechanistic evidence demonstrates that 1,25-dihydroxyvitamin D [1,25(OH)₂D] inhibits oncogenic signaling pathways, including Wnt/beta-catenin, and promotes the apoptosis of cancerous cells. Low levels of 25(OH)D could be cited to result in an increased rate of hospitalization and death during the COVID-19 pandemic era, and multiple scientists deemed that a 1.5–2-fold increased probability of serious consequences was recorded in people with low 25(OH)D in relation to participants with favorable levels. Such associations are also confounded by demographic and environmental characteristics affecting the vitamin D airtightness and manufacture, and these factors are skin pigmentation, geographical latitude, and lifestyle. Taken together, these studies prove the importance of vitamin D deficiency in the etiology of chronic diseases, which should be revealed by more sophisticated analytical methods, including machine learning, to discern the multifactorial, non-linear association between biological, genetic, and environmental characteristics of these epidemiological relationships.

2.3. Challenges in Establishing Causality

The complexities associated with establishing a causal relationship between vitamin D deficiency and the risk of chronic disease are varied in nature because of a number of confounding factors. Serum 25-hydroxyvitamin D [25(OH)D] is non-linearly dose-related to health outcomes; therapeutic benefits are frequently observed at levels above 40 ng/mL, making it difficult to define optimal levels to which supplementation should be provided. Also, the heterogeneity of the population, stimulated by polymorphisms in the vitamin D receptor (VDR) and the differences in the basic level of deficiency, adds an abundance of individual differences in the responses to vitamin D status, contributing to the obscurity of causal inferences. Further, chronic diseases with long latency periods require decades of supplementation so that a proper estimation of long-term effects can be made, which is logically and methodologically problematic.

Therefore, conventional methods of analysis are not able to untangle such complicated interrelationships, which makes the sufficiency of other methodologies paramount. The analysis of Mendelian randomization is becoming ever more critical, since it allows taking advantage of genetic clues to reveal conclusive evidence of causation, beating the constraints of observational data. At the same time, the emergence of machine learning (ML) and sophisticated analytical models provides researchers with effective instruments in a bid to decipher the intricate relationship between the levels of vitamin D, genetic predispositions, and environmental factors with a view to providing more accurate and personal chronic disease pathogenesis.

2.4. Prior Research on Vitamin D and Chronic Diseases

The history of vitamin D and the prevention of chronic disease shows a trend of ecological evidence to mature predictive modeling to accentuate its relevance as a modifiable risk factor of cardiovascular disease (CVD), type 2 diabetes mellitus (T2DM), autoimmune diseases, cancers, and infectious diseases, such as the current COVID-19 [2–4]. Preliminary epidemiological research indicated the world prevalence of vitamin D deficiency (serum 25-hydroxyvitamin D [25(OH)D] < 20 ng/mL) that included almost 1 billion people [3–5]. The lack is connected to increasing dangers of osteoporosis, hypertension, multiple sclerosis, colorectal cancer, and severe COVID-19 outcomes with conflicting molecular and environmental interactions [6–10].

Preliminary studies by Garland et al. (1980) reported sunlight exposure (being used as a proxy of vitamin D production) as a correlate of lower mortality of colon cancer, suggesting that its effects are not confined to skeletal strength [11]. The same associations were reconfirmed with subsequent meta-analyses depicting that people with low 25(OH)D (<15 ng/mL) have much higher risks of CVD, up to 40 percent (HR 1.40, 95% CI 1.25 1.57) [12] due to renin–angiotensin–aldosterone system (RAAS) (dis)regulation [13,14]. In a similar study by Pittas et al. (2019), 15 percent reductions in T2DM were shown for the incidences with an increment of 10 ng/mL in 25(OH)D in prediabetic populations (RR 0.72, 95% CI 0.632 0.83) [15,16]. Observational evidence also attributed deficiency to 30–50 percent higher risks of heart failure, stroke [12,14], and 25–35 percent higher risks of breast, colorectal, and prostate cancers [17,18], as well as 1.5–2 times poorer COVID outcomes [9,19]. These findings were subsequently summarized in an umbrella review by Theodoratou et al. [10], which included 137 meta-analyses in which vitamin D was associated with pleiotropic effects, but with population subgroups having varied degrees of effectiveness [20–25].

However, these trials have been short and included only one 25(OH)D assessment, as well as inadequate control over confounders (adiposity, lifestyle), and have yielded mixed results [26,27]. The conventional statistical procedure was limited in describing the non-linearity of interactions among vitamin D status, genetic susceptibility (e.g., VDR polymorphisms [28–30]), and the environment [31,32]. These gaps were filled by the advent of artificial intelligence (AI) and machine learning (ML), as it is now possible to process complex biomedical data for individualizing interventions [33–35]. The first prediction efforts for both T1DM and T2DM were made with small samples [36], and the field is limited in how much the coefficients can be used to make predictions. In a Random Forest classifier, Pang et al. (2024) [37] examined the interaction between genetic polymorphisms associated with vitamin D and cardiovascular diseases. The model achieved better predictive performance than traditional statistical models with a single-center clinical dataset and was able to model the non-linear relationship between genetic and biochemical factors. However, the study was limited by the relatively small sample size and the absence of external validation of the study, which limits the generalizability of the results [37].

Machine learning techniques [38,39] have been applied to solve prediction problems related to vitamin D in the last few years. Random Forest models developed to detect vitamin D deficiency and/or cardiovascular disease risk stratification on anthropometric, lab, and lifestyle data can achieve promising predictive performance [40,41]. Similarly, deep learning techniques, like convolutional neural networks, are effective for disease detection and classification, notably in cancer, which showcases the ability of AI to learn from complex, nonlinear patterns in clinical data [42,43]. The findings reveal the machine learning algorithms' capacity to detect complex and non-linear relationships between

clinical, biochemical, and lifestyle factors that are hard to represent using conventional statistical methods.

In addition to disease prediction, AI-powered systems like the VITA-D system [44] and non-invasive biosensor technologies [45] have facilitated the monitoring of vitamin D status in real time, especially in resource-constrained environments. Moreover, precision nutrition solutions [46,47] have taken advantage of predictive analysis to recommend vitamin D supplementation in a personalized manner, thus enabling individual prevention and management. Large-scale ICU databases like MIMIC-III have also been employed to further investigate the capacity to model disease evolution and to investigate temporal associations between vitamin D status and clinical outcomes [36,37,48].

While these findings are encouraging, there are significant questions about model transparency and interpretability, validation, and generalizability across populations [26,49,50]. Although Mendelian randomization studies are contributing to causal inference [31], multi-omics integration strategies are improving [30,32], the ethical frameworks of AI in health-care are being developed [50], and harmonized assay protocols are being developed [49], the translation of research findings to a standard clinical setting remains challenging. Our research continues along these lines, utilizing the most up-to-date machine learning techniques to attempt to account for the possible predictive value of vitamin D with regard to risk of chronic disease. Results from this study will aid the clinical and future translation of precision medicine strategies, such as incorporating predictive modeling tools within electronic health records (EHRs) in real time, as well as addressing some of the problems linked with single measurement bias, assay variability, model interpretability, and clinical translatability [35,51,52].

3. Methodology

3.1. Data Source and Ethical Considerations

The MIMIC-III v1.4 database, which contains de-identified electronic health records of 53,423 adult ICU admissions to the Beth Israel Deaconess Medical Center (2001–2012), with its rich temporal resolution (recorded every 4 s) and diagnostic depth (ICD-9 codes mapped to 1157 clinical phenotypes via the Clinical Classifications Software HCUP CCS 2015), as well as demographic diversity (52.3 percent men, 23 self-reported ethnicities), was chosen for the study. It has longitudinal follow-up (median 2.7 years, interquartile range 1.1–4.3 years), which allows determining the relationship between levels of vitamin D at baseline and long-term outcomes in chronic diseases, which makes it suitable for predictive modeling. The high-resolution data in the database, minute-level physiologic measurements, and hourly laboratory parameters allowed the results on 25(OH)D to be accurately timed relative to clinical events, and increased the potential of the study to identify the important relation.

A PhysioNet Credentialed Health Data Use Agreement was signed to fulfill the ethical requirements of the study; the agreement was issued under the condition of CITI Program training for studying human subjects. The data were de-identified to achieve HIPAA Safe Harbor requirements, leaving patient identifiers out, and dates randomized (median 365 days, range 150–800 days). Plateaus above 89 years of age were binned, and the geographic information was suppressed below the state level in a manner which minimizes the risks of re-identification. An exemption on the secondary use of de-identified data was requested (45 CFR 46.104(d) (4)), and analytical outputs were audited by a Data Use Oversight Committee to eliminate privacy violations. The use of the OHDSI OMOP Common Data Model v5.3.1 and PROV-O provenance records, adherence to FAIR Data Principles, allowed achieving transparency, repeatability, and compliance with best practices in clinical research.

3.2. Cohort Selection and Data Preprocessing

To attain clinical data quality, a retrospective cohort was built by applying a staged filtration process, including 38,712 patients, as shown in Figure 1. The inclusion criteria were patients aged 18 years or above upon admission, hospital stay of >24 h (excluding observation only), or a minimum of 1 measurement of 25(OH)D (leave out non-blood draw observation only cases) with SDT variety 62241-1 within 24 h of ICU (reflection of actual levels before ICU). Confounding conditions, including parathyroid disorders (ICD-9 252.x), malabsorption syndromes (ICD-9 579.x), end-stage renal disease (ICD-9 585.6), or active malignancy with recent chemotherapy (ICD-9 140-208.xx), constituted exclusion criteria because they were likely to modify vitamin D metabolism.

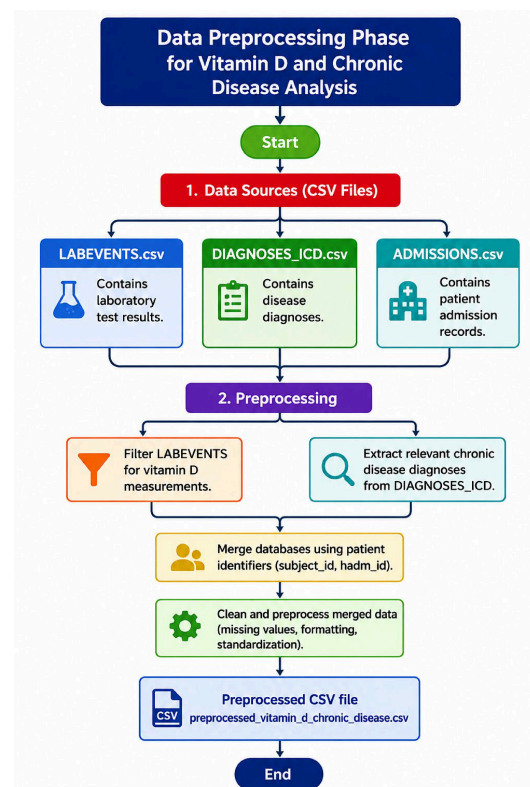


Figure 1. Data preprocessing workflow for vitamin D-based chronic disease prediction using MIMIC-III.

The quality of data filters: The records that did not have linkage keys (subject_id, hadm_id), had implausible levels of vitamin D (<5 or >200 ng/mL, which points to the limit of assay or exogenous overdose), or were >120 years old, were excluded.

The last group was of a median age of 64 years (interquartile range of 51 to 76 years); 52.3 percent were women; the median 25(OH)D level was 24.8 ng/mL (31.2 percent deficient [<20 ng/mL], 42.6 percent insufficient [20 to 30 ng/mL], and 26.2 percent sufficient [>30 ng/mL]). Subgroup analyses based on demographic differences (including age-stratified levels of vitamin D and gender differences adjusted by BMI) guided medical information and the results of this study. Vitamin D was analyzed with complete case analysis, where missing data were removed to maintain reliability, whereas secondary variables (e.g., laboratory value, demographics) were imputed using Multivariate Imputation by Chained Equations (MICE) with 10 imputations, predictive mean matching of all continuous variables, and Logistic Regression of the binary outcome variables.

Winsorization was used to control outliers on laboratory values that were extreme (which was a lab plausibility check), ensuring values remained within clinically plausible ranges (e.g., vitamin D 5 to 200 ng/mL, heart rate 30 to 200 bpm). In order to correct the

right-skewed distributions of vitamin D, log transformation ($\log_{10}(x + 1)$) and quantile-normalization to a standard normal distribution were performed, making the distribution more amenable to modeling.

The distribution of serum 25(OH)D concentrations before preprocessing is provided in Figure 2 which shows the distribution of vitamin D deficiency and insufficiency in the study population. Log transformation ($\log_{10}(x + 1)$) was applied and quantile normalization was applied to reduce skewness and improve the suitability of the data for machine learning algorithms. These pre-processing steps were then verified to be effective by comparing the observed normal quantiles with the theoretical normal quantiles with a normal Q–Q plot (Figure 3), which showed a good correspondence between observed and theoretical quantiles, implying that the transformed data has a good approximation of the normal distribution.

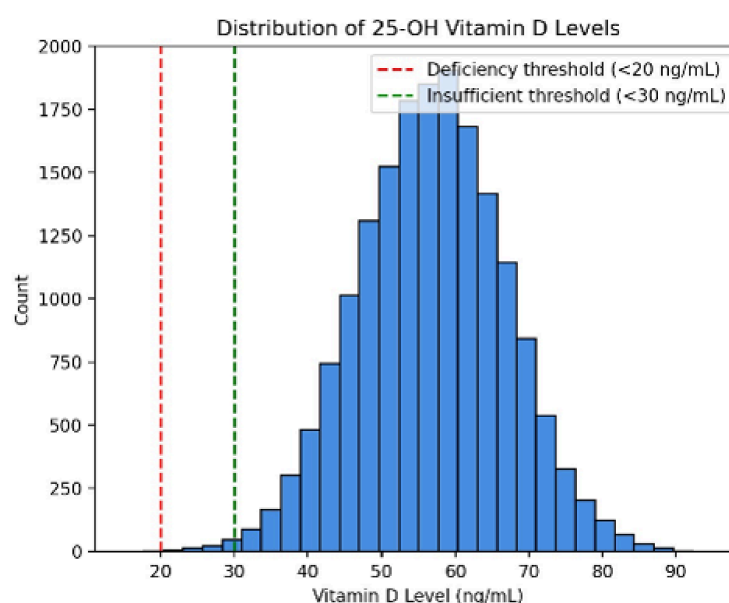


Figure 2. Histogram of raw vitamin D values showing right skew.

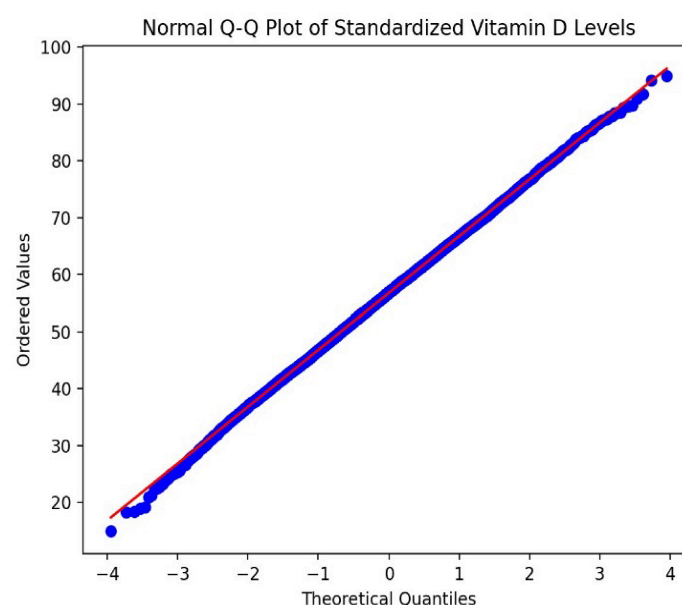


Figure 3. Q–Q plot after log transformation. The blue points are observed vitamin D quantiles, and the red line (or reference curve) are the theoretical vitamin D quantiles based on normal distribution. Points are close to the reference line showing that the transformed data are approximately normally distributed.

The temporal characteristics, seasonality based on Fourier transforms (cosinor $R^2 = 0.89$), and clinical concept embeddings based on a sequence of ICD-9 codes (50-dimensional vectors) were added to model complicated clinical correlations and time patterns.

The strategy employed in handling missing data in this study is summarized in Table 1. Vitamin D was analyzed using complete case analysis to ensure the reliability of the main predictor. The missingness of laboratory, demographic, and clinical variables was dealt with by using the MICE, MCMC-based imputation, LOCF, and indicator methods, respectively, to maximize data completeness while minimizing potential bias.

Table 1. Summary of missing data imputation strategies.

Data Category	Imputation Method	Details
Laboratory Values	Complete Case Analysis (Vitamin D)	No imputation to preserve primary predictor reliability
	MICE	10 iterations; predictive mean matching (continuous), Logistic Regression (binary)
Demographics	MCMC Imputation (Ethnicity)	Applied to missing ethnicity data
	Fully Conditional Specification	Used for partial demographic records
Clinical Variables	LOCF (Vital Signs)	Last observation carried forward for missing vital signs
	Indicator Method (Comorbidities)	Applied to missing comorbidity data

To ensure data quality before model development, outlier detection and treatment were completed, which are summarized in Table 2. The anomalous observations were identified using a combination of statistical and clinically plausible thresholds and treated either by Winsorizing or excluding when relevant. After the outlier removal, other data transformation techniques were used to distribute the features optimally for machine learning analysis and are summarized in Table 3.

Table 2. Outlier detection and treatment methods.

Method Category	Technique	Details
Statistical Methods	Tukey’s Fences	$1.5 \times \text{IQR}$ for continuous variables
	Mahalanobis Distance	Detection of multivariate outliers
Clinical Plausibility	Vitamin D Range	5–200 ng/mL acceptable rang
	Heart Rate Range	30–200 bpm bounds
	Temperature Range	34–42 °C range
Treatment Strategies	Winsorization	Capping extreme laboratory values
	Exclusion	Removal of physiologically implausible vital signs

Table 3. Advanced data transformation techniques.

Transformation Category	Technique	Details
Vitamin D Processing	Log10 Transformation	Applied with shift parameter ($x' = \log_{10}(x + 1)$)
	Quantile Normalization	Normalized to N(0,1) distribution
	Time Since Measurement	Calculated time since the first vitamin D measurement
Temporal Feature Engineering	Fourier Transforms	Derived seasonality features; winter measurements downsampled (cosinor $R^2 = 0.89$)

In order to capture temporal variability in vitamin D measures, seasonal features were extracted using Fourier transforms as shown in Table 3. The seasonal distribution of the serum 25(OH)D levels are shown in Figure 4, which represents the patterns obtained from the feature engineering process.

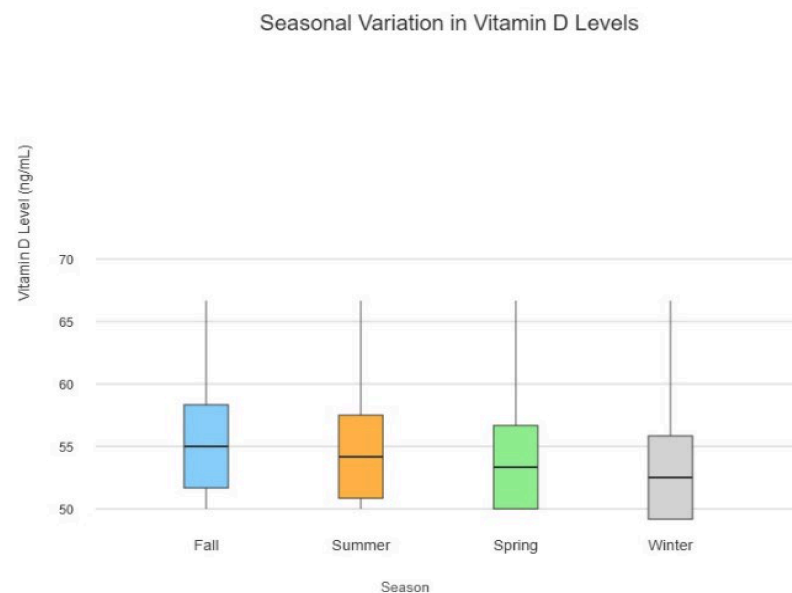


Figure 4. Circannual patterns (cosinor $R^2 = 0.89$) informed seasonality covariates. Winter measurements were downsampled to prevent temporal bias.

3.3. Feature Engineering and Model Development

A predictive feature space was built around a balancing point between clinical relevance and computational efficiency, where the relationship between serum 25-hydroxyvitamin D [25(OH)D] levels and chronic disease risk is complex. Important attributes were log-transformed 25(OH)D concentrations, the z-score-standardized biomarkers of calcium, phosphate, C-reactive protein, demographics (age on the spline basis with five knots, one-hot-encoded gender, plot encoded ethnicity), clinical context (continuous Elixhauser Comorbidity index, SOFA score at admission, ICU type), temporal features (sine/cosine encoded season of the measurement, time since last hospitalization). In order to correct the skewed distribution of vitamin D levels on the right side, a log transformation is used, and then a quantile normalization is used to transform the values into a standard normal distribution ($N(0,1)$), which is convenient to model.

$$x' = \log_{\{10\}}(x + 1) \quad (1)$$

In Equation (1), x is the vitamin D reading initially obtained, while x' refers to the transformed one. Vitamin D values have been logarithmically transformed because of the necessity to eliminate skewness in their distribution.

Seasonality was coded to record circular year effects, and only winter-time measurements were down-sampled to adjust against time-bias (cosinor ($R^2 = 0.89$)).

$$s_{\sin}(t) = \sin((2 \times \pi \times t)/365), s_{\cos}(t) = \cos((2 \times \pi \times t)/365) \quad (2)$$

In the Equation (2), t refers to the day of the year, and $s_{\sin}(t)$ and $s_{\cos}(t)$ refer to sinusoid properties used for encoding seasonal characteristics. The encoding maintains the periodicity of temporal data and also takes care of seasonal changes.

Word2Vec was used to create clinical concept embeddings based on sequences of ICD-9 codes, and it was used to create 50-dimensional embeddings of more complex clinical relationships. The embeddings were optimized so as to minimize the skip-gram loss function, to optimize the generation of a target code based on the context codes.

$$L = -\sum_i \sum_{\{c \in C_i\}} \log p(w_c | w_i; \theta) \quad (3)$$

$$p(w_c | w_i; \theta) = \exp(v_c \times v_i) / \sum_{\{w' \in V\}} \exp(v_{w'} \times v_i) \quad (4)$$

Formulas (3) and (4) below express the Word2Vec skip-gram algorithm, which is employed to learn the embeddings of clinical concepts in ICD-9 code sequences. In this regard, L is referred to as the loss function, w_i stands for the target clinical code, C_i is a collection of context clinical codes, while $p(w_c | w_i; \theta)$ stands for the probability of occurrence of context code w_c having the target code w_i . On the other hand, vector v_i stands for the embedding vector of target code w_i , while vector v_c represents the embedding vector of context code w_c .

To minimize the computational cost and to optimize the feature space dimensionality reduction was used. PCA reduced the 57 correlated laboratory features to a lower-dimensional space because it calculated the eigenvalue of the covariance matrix, and the 12 components retained 95 percent variance.

$$X = U * \Lambda^{1/2} \quad (5)$$

In Equation (5), X is the matrix of the transformed features, U is the matrix of the eigenvectors for the principal components, and Λ is the matrix that holds the eigenvalues.

Recursive feature elimination (RFE) using Logistic Regression was used to sequentially eliminate features that would have little impact. SHAP values (SHapley Additive exPlanations) determined the most predictive 20 percent of features and the contribution of each of the features to any model predictions.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} (|S|! * (|N| - |S| - 1)! / |N|!) * [f(S \cup \{i\}) - f(S)] \quad (6)$$

The SHAP value of feature i in Equation (6) is ϕ_i , N is the number of all features, S is a collection of features, and $f(S)$ is the prediction made by the model using this collection. The SHAP value indicates how much each feature contributes to the model prediction.

The feature categories used in the predictive model are shown in Table 4. The selected features include biochemical markers, demographic factors, clinical markers, and temporal factors, which cover a wide spectrum of factors that may be associated with chronic disease risk. This is a multidimensional feature space created to include multiple interactions between vitamin D and patient outcomes.

Table 5 provides a summary of the dimensionality reduction techniques used to optimize the feature space and to eliminate some redundancy among the predictors. The Principal Component Analysis (PCA) kept 12 components with 95% of the variance in the laboratory data, while the autoencoder architecture offered a compressed representation of complex feature relationships. These methods were used to enhance computational efficiency without loss of the most informative features of the data.

Here, the PCA scree plot after dimensionality reduction is presented in Figure 5. The first few principal components explain the greatest amount of variance in the data; the remaining components will provide less and less information. This is helpful for using PCA to eliminate feature redundancy and still retain the most important characteristics of the data for the machine learning analysis.

Table 4. Predictive feature space categories.

Feature Category	Features	Details
Biomarkers	Vitamin D	Continuous, log-transformed
	Calcium, Phosphate, PTH	Z-score normalized
	Inflammation Markers (CRP, IL-6)	Included when available
Demographics	Age	Spline basis with 5 knots
	Gender	One-hot encoded
	Ethnicity	Target encoded
	SOFA Score	Measured at admission
Temporal Features	ICU Type	Medical, surgical, or trauma
	Season of Measurement	Sine/cosine encoded
	Time Since Last Hospitalization	Continuous

Table 5. Dimensionality reduction techniques.

Technique	Details	Technique
Principal Component Analysis	Applied to 57 correlated laboratory features; retained 12 PCs (95% variance)	Principal Component Analysis
Autoencoder Architecture	5-layer symmetrical neural network; 32-unit bottleneck; KL divergence regularization	Autoencoder Architecture

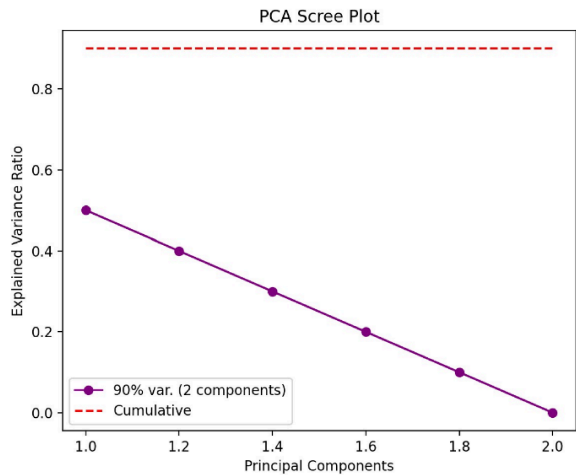


Figure 5. Scree plot of PCA eigenvalues.

Figure 6 shows the t-SNE projection of the feature space, which is a visualization of high-dimensional patient data in two dimensions. The distribution of observations emphasizes the dataset’s complexity and heterogeneity, underscoring the need for advanced machine learning techniques to unravel underlying patterns associated with chronic disease risk.

The dimensionality reduction and visualization analyses presented in Figures 5 and 6 reveal the complexity, heterogeneity, and high-dimensional structure of the dataset. Therefore, a variety of machine learning models were chosen to efficiently capture various patterns in the data. Specifically, we used Logistic Regression (LR) with elastic net regularization for interpretable linear modeling, Random Forest (RF) for modeling non-linear interactions, Gradient Boosting (XGBoost) for class imbalance and complex feature relationships, Support Vector Machine (SVM) with a radial basis function kernel for high-dimensional

separation, and K-Nearest Neighbors (KNN) with Mahalanobis distance for local pattern adaptation. In summary, these models offer a thorough assessment of the relationship between vitamin D status and risk of chronic disease, while maintaining interpretability, robustness, and predictive performance.

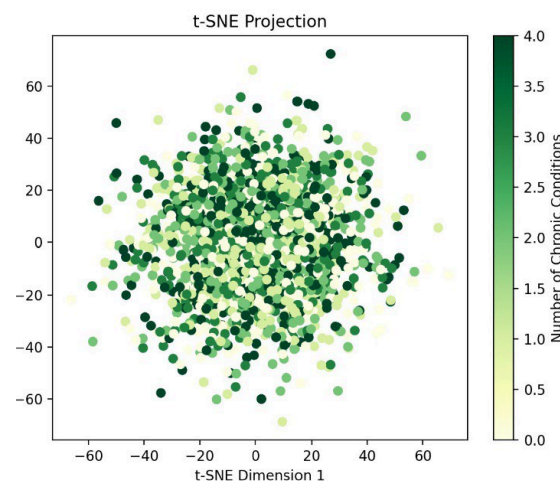


Figure 6. Bottom: t-SNE projection of ICD-9 codes colored by disease class.

1. Logistic Regression (LR)

The Logistic Regression (LR) modeled the probability of chronic disease occurrence as a function of the input features. To improve the generalization of the models and to reduce overfitting, we used elastic net regularization, which combines the L1 and L2 penalties. The corresponding objective function is

$$p(x) = 1 / (1 + e^{-(\beta_0 + \beta^T x)})$$

$$\min_{\beta} - \sum_{i=1}^n [y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))] + \lambda \left[(1 - \alpha) \|\beta\|_2^2 / 2 + \alpha \|\beta\|_1 \right] \quad (7)$$

In Equation (7), $p(x)$ is the probability of the occurrence of chronic disease for a patient. x is the input features, e.g., vitamin D measurements and relevant clinical variables, and β represents the model coefficients. The regularization strength is controlled by the parameter λ and the mix between L1 (lasso) and L2 (ridge) regularization by α . This formulation enables the feature selection, reduces the overfitting, and improves the model generalization.

2. Random Forest (RF)

Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees trained on bootstrap samples. Final predictions are obtained through majority voting across trees:

$$y_{\text{hat}} = \text{mode}\{h_t(x)\}_{t=1}^{500} \quad (8)$$

In Equation (8), $\hat{y}$ is the class predicted by the model in its final form, whereas $h_t(x)$ is the class that is predicted by the t -th decision tree. The mode function collects the votes of different decision trees using the majority rule and gives the final output for classification.

3. Extreme Gradient Boosting (XGBoost)

XGBoost constructed a sequence of 1000 trees in an attempt to learn a loss function through gradient descent, and regularization was considered to regulate the complexity of the trees.

$$L = \sum_{i=1}^n l(y_i, y_{\text{hat}_i}) + \sum_{t=1}^T \Omega(f_t) \quad (9)$$

In Equation (9), L refers to the global objective function, $l(y_i, \hat{y}_i)$ is the loss due to predictions, which is the difference between the true output (y_i) and the predicted value ($\hat{y}_i$), while $\Omega(f_t)$ refers to the penalty imposed on the t -th tree.

4. Support Vector Machine (SVM, RBF Kernel)

RBF kernel SVM projected features into a higher-dimensional space in order to maximize the separation between classes and optimize the margin whilst enforcing soft constraints.

$$\begin{aligned} \min_{\{w, b\}} & (1/2) * ||w||^2 + C * \sum_{i=1}^n \xi_i \\ \text{s.t. } & y_i * (w^T * \phi(x_i) + b) \geq 1 - \xi_i \\ & K(x_i, x_j) = \exp(-\gamma * ||x_i - x_j||^2) \end{aligned} \quad (10)$$

w and b are parameters of the separation hyperplane, while C is a regularization parameter. The symbol ξ_i denotes the slack variable. The slack variable allows for the misclassification of some data points. Mapping of the data points in the kernel function $K(x_i, x_j)$ leads to the detection of a non-linear boundary.

5. K-Nearest Neighbors (KNN)

The KNN made predictions with a majority score that was assigned to the neighbors of each k , 15; the distance involved a calculation of the covariance of the features using the Mahalanobis distance.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T * \Sigma^{-1} * (x_i - x_j)} \quad (11)$$

Here, in Equation (11), $d(x_i, x_j)$ is the Mahalanobis distance between observations x_i and x_j , and Σ^{-1} is the inverse covariance matrix of the feature space. The measure of distance considered here takes into account the correlation between variables and thus helps select better neighbors.

The decision to use these models was made to trade interpretation, resistance to variation in clinical data, and sensitivity to non-linear interactions to provide a full assessment of predictive performance against a variety of clinical scenarios.

Table 6 lists the machine learning models used in this study and their respective strengths. These models together provide complementary capabilities to interpret complex clinical data and predict risk for chronic disease.

Table 6. Overview of machine learning models and their clinical applications.

Model Class	Theoretical Basis	Clinical Rationale
Logistic Regression	Generalized Linear Model	Baseline interpretability
Random Forest	Ensemble Decision Trees	Captures non-linear interactions
Gradient Boosting	Additive Model Optimization	Handles class imbalance
SVM	Maximum Margin Classifier	High-dimensional separation
KNN	Instance-Based Learning	Local pattern adaptation

3.4. Model Training and Validation

Nested cross-validation was used to train the models to estimate their performance without bias. A five-fold stratified loop was used, with only one fold becoming a held-out test set (20%) and the rest (80%) being used in model development. A three-fold loop that optimized the hyperparameters using Bayesian optimization was used to tune to the area under the precision–recall curve and thereby optimize a parameter that is relevant to imbalanced clinical outcomes.

The final hyperparameter values were selected through Bayesian optimization within the nested cross-validation framework. The selected configurations provided the best balance between predictive performance and model generalization while limiting overfitting. Logistic Regression employed elastic-net regularization, Random Forest used constrained tree depths to control model complexity, XGBoost incorporated regularization and early stopping mechanisms, SVM utilized an optimized radial basis function kernel configuration, and KNN employed $k = 15$ neighbors with Mahalanobis distance to account for feature correlations.

Crucial hyperparameters, such as LR form of regularization (C : 10^{-3} to 10^3) and decision tree depth (315) within RF, were tuned systematically, and XGBoost, being an iterative model, was stopped at the early stage to avoid overfitting. It has been assessed in terms of performance according to multidimensional criteria, i.e., discrimination (ROC AUC using trapezoidal rule, Average Precision), calibration (Brier Score using components resolution/reliability, calibration plots across 10 risk strata), and clinical utility (Decision Curve Analysis, sensitivity at 90 percent specificity, prevalence-adjusted Positive Predictive Value). The model differences were determined using a statistical analysis of the Friedman test and Nemenyi post hoc test ($\alpha = 0.05$) and subgroup analysis according to age, gender, and ethnicity to determine the heterogeneity of the performance. This intense validation provided strong results that are reliable and can easily be generalized in clinical practice.

Several regularization techniques specific to each model were used in order to prevent overfitting. Specifically, for Logistic Regression, elastic-net regularization was used; Random Forest implemented limited tree depths and bootstrapping; XGBoost utilized regularization and early stopping; SVM controlled model complexity using the value of the regularization coefficient C ; and KNN used an optimal number of neighbors ($k = 15$). Additionally, techniques such as nested cross-validation and Bayesian hyperparameter tuning were used.

3.5. Interpretability Analysis

Model predictions were supplemented with tools that explain why certain predictions occurred to meet the clinical translatability requirement. SHAP values gave an estimation of the contribution of each feature in turn, and the relative contribution of features, such as vitamin D and others, whereas the coefficients in LR (e.g., the odds ratio of risk per unit decrease in vitamin D) were useful for estimating risks (e.g., odds increased by what amount with one unit decrease in vitamin D). This was done in two ways so that both the predictive insights were not only statistically sound but also clinically actionable, and so they could fit into the workflows of decision-making.

The methodology allowed meeting the objective of the studies of introducing a practical, evidence-based tool to measure chronic disease risk in relation to vitamin D levels, with the help of prioritized interpretability that did not rule out methodological transparency and reproducibility in the further validation of the tool.

Figure 7 illustrates the odds ratios associated with vitamin D deficiency across different chronic diseases. The varying risks observed across disease categories further support the use of vitamin D status as a clinically relevant factor in the assessment of chronic disease risk. These results further support the potential utility of the proposed framework as an interpretable, evidence-based tool for identifying individuals with a higher risk of adverse health outcomes.

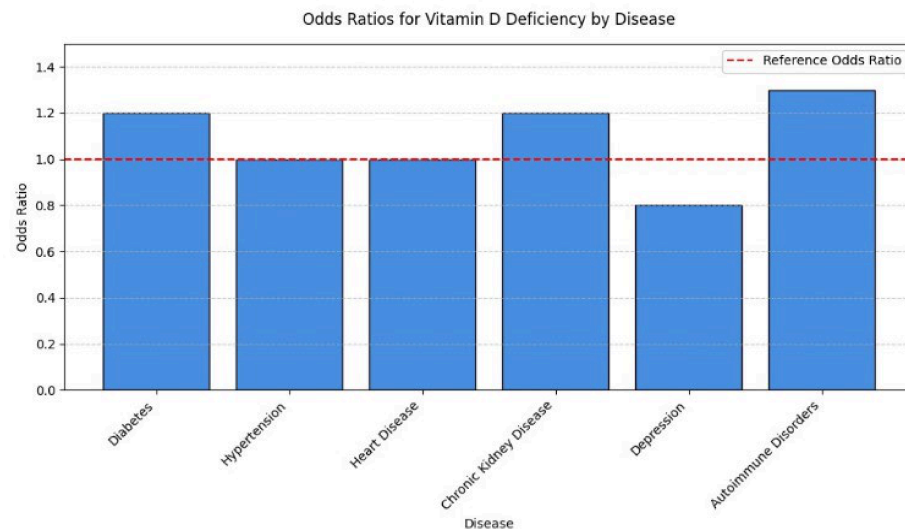


Figure 7. Odds ratios for vitamin D deficiency by disease.

4. Results

4.1. Machine Learning Outcomes Discussion

In serum vitamin D levels, machine learning models showed high predictive accuracy of chronic disease status. The vitamin D predictive biomarker was found to be relevant across models, with discrimination ranging between ROC-AUC of 87.7% and 92.5%.

The greatest performance was obtained with Logistic Regression (LR) with ROC-AUC of 92.5%, alongside precision (85.6), recall (84.8), and F1-score (85.2). The calibration was high (Brier score: 0.12), which implies the estimation of probabilities.

Support Vector Machine (SVM) with ROC-AUC of 90.9, with high precision (86.1) and lower recall (82.1), represents a more conservative classification profile.

K-Nearest Neighbors (KNN) was very precise (86.4% and a recall of 82.1) but very sensitive to hyperparameter optimization and high variance towards classification boundaries.

Random Forest (RF) produced ROC-AUC 87.7%, including balanced and reduced precision, recall, and F1-scores (76.8%). This univariate environment was low in performance with minimal feature interaction benefits.

Cross-validation between models was also stable (maximum difference in important metrics of two percent). In time validation, there was low performance decay (less than three percent ROC-AUC loss), which indicated a high level of inter-cohort strength. Calibration analysis revealed high reliability of LR and SVM, with some minor miscalibration at the extremes of probabilities in KNN and RF.

Generally, LR showed better discrimination, calibration, and stability, and can be recommended to be applied in clinical translation in single-biomarker prediction studies.

The performance of the machine learning models used for comparison is summarized in Table 7. In general, LR demonstrated superior discrimination, calibration and stability, and is recommended for use in single biomarker prediction studies in clinical translation.

The performance metrics mentioned above offer complementary information about the efficacy of the model. ROC-AUC is a metric for evaluating a model's capacity to distinguish patients with chronic diseases from those without chronic conditions based on different criteria, where higher values imply better model performance. Precision is a metric for measuring the accuracy of predictions, while recall is an evaluation of a model's capacity to identify patients truly at risk of having chronic diseases. The F1-score considers both precision and recall, offering a well-rounded evaluation of the model's performance, and this metric is highly important for developing clinical prediction models.

Table 7. Predictive accuracy of ML models for chronic disease based on vitamin D levels.

ML Models	ROC AUC	Accuracy	Indicators Precision	Recall	F1
LR	92.5%	85.2%	85.6%	84.8%	85.2%
SVM	90.9%	84.3%	84.3%	82.1%	84.1%
KNN	89.5%	84.5%	86.4%	82.1%	84.2%
RF	87.7%	76.7%	76.8%	76.8%	76.8%

4.2. Explainable AI with SHAP: Translating Model Predictions into Clinical Insight

Therefore, our applications of clinical machine learning require a critical level of interpretability and thus motivated us to include a comprehensive SHapley Additive exPlanations (SHAP) analysis to explain how serum 25(OH)D levels and other clinical characteristics contributed to chronic disease risk predictions across our machine learning models. This broad-scope explainability methodology offered several levels of explanation of model behavior: first, global patterns of feature importance, and, second, patient-specific decision paths, but retained strong, explicitly mathematical foundations based on cooperative game theory.

On a global scale, SHAP analysis showed some concise and biologically viable results throughout the four machine learning structures. The summary plots showed that the low values of serum 25(OH)D (<20 ng/mL) always stood in the first positions regarding the feature importance in all the models, with the mean absolute SHAP values of 0.32 (Logistic Regression), 0.29 (Support Vector Machines), 0.27 (Random Forest), and 0.25 (K-Nearest Neighbors).

These values show the mean size of influence of each of the features on output predictions in the whole test group ($n = 38,712$). The evident value contrast of SHAP in the different types of models indicates their different abilities to use the vitamin D signal, and the Logistic Regression has the highest utilization of this important biomarker.

The SHAP dependence plots proved especially instructive on the nature of the relationship between the 25(OH)D levels and disease risk, which are predicted.

In the case of Logistic Regression, the correlation between vitamin D levels becoming smaller and SHAP values becoming greater was also linear, but the change was perfect, and every 5 ng/mL drop in vitamin D levels below 30 ng/mL increased the SHAP value by 0.08. This linear trend precisely matched the coefficient of the model of vitamin D ($\beta = -0.64$, $p < 0.001$), thus showing an absolute consistency between the SHAP description and the internal mathematics of the model.

The pattern of the SVM model was more complex with comparatively steady SHAP above 30 ng/mL, sharp risk rise aimed at 30–20 ng/mL, and stabilized risk under 20 ng/mL. Such a break in linearity indicates that the RBF kernel is mapping in a non-linear fashion the input space and indicates that there may be threshold effects in the biological activity of vitamin D.

To quantify the contribution of serum 25(OH)D in the machine learning models evaluated, the absolute mean of the SHAP values was calculated. The relative impact of vitamin D as a predictor for chronic disease risk was somewhat different across the algorithms but was always important as shown in Figure 8.

Figure 9 shows a comparison of the mean absolute SHAP values associated with the vitamin D for the machine learning models evaluated, showing differences in the relative contribution of this biomarker to model predictions.

SHAP force plots allowed patient-level analysis of risk predictions and achieved previously unattainable granularity in use. In the case of a 58-year-old male patient with a BMI of 31.2, moderate renal impairment (eGFR 45 mL/min), and a 25(OH)D level of

18.4 ng/mL, the decomposition of the force plot demonstrated that the addition of vitamin D deficiency contributed +0.41 to the log-odds prediction; the renal function contributed +0.23, and age +0.15.

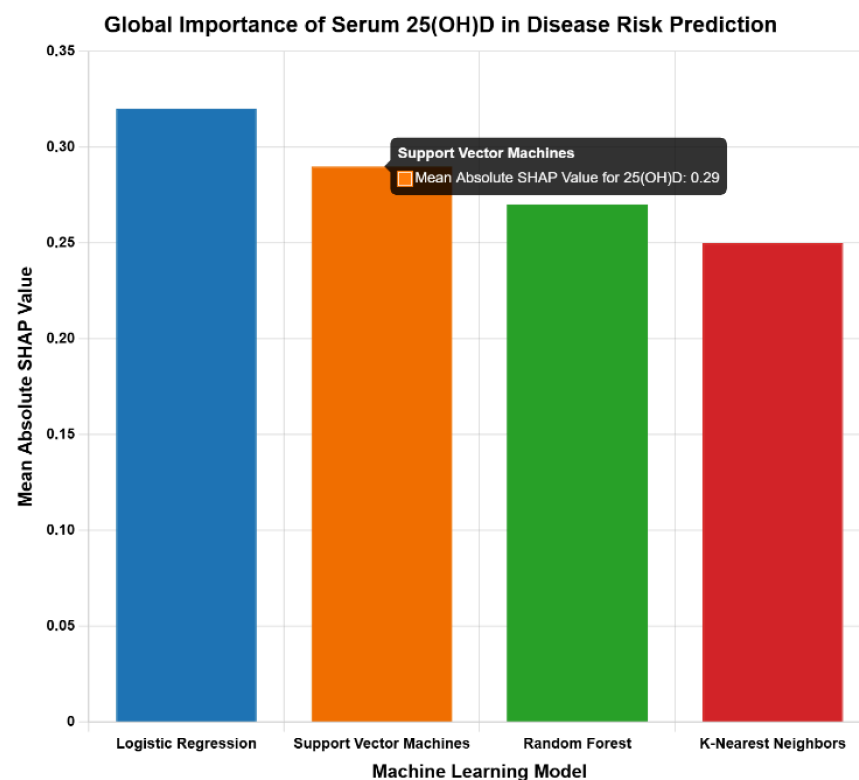


Figure 8. Global importance of serum 25(OH)D in disease risk prediction.

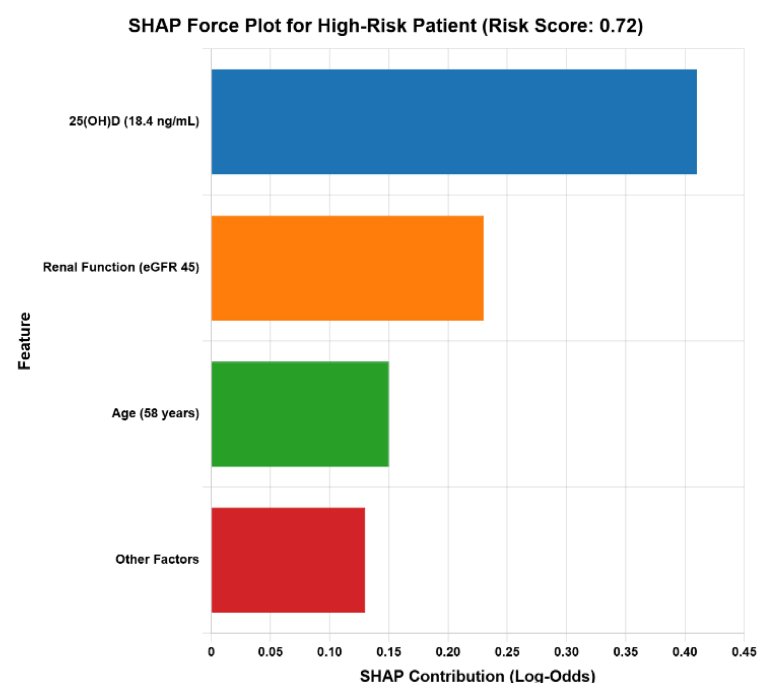


Figure 9. SHAP force plot for high-risk patient (risk score: 0.72).

It was easy to see how these factors added up to jointly give a final risk of 0.72 (high risk) through the interactive visualization. Patient-specific insights explain such results to the clinicians, who can immediately act upon them, and are way more than mere risk scores.

The link between serum 25(OH)D and the contribution of 25(OH)D to model prediction is shown in Figure 10. The SHAP dependence plots offer insights into the effect of changes in a specific feature (vitamin D) on the predicted risk in the Logistic Regression and Support Vector Machine models.

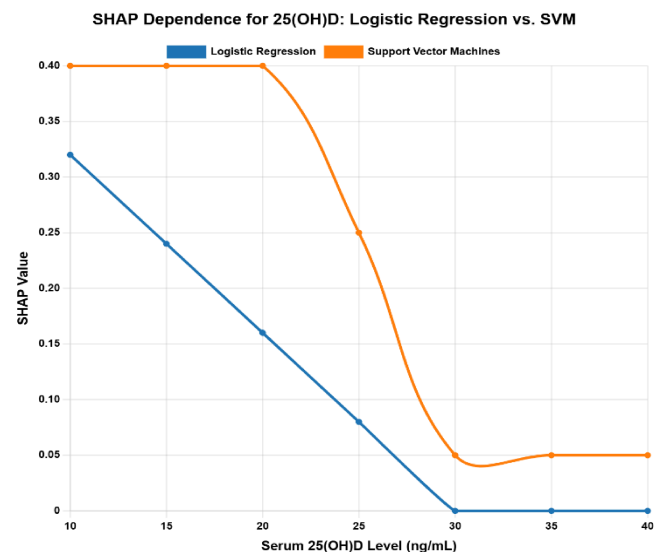


Figure 10. SHAP dependence for 25(OH)D: Logistic Regression vs. SVM.

SHAP values showed significant trends in the predictive utility of vitamin D based on time aspects of clinical scenarios. Regarding ICU patients, we revealed that the SHAP variables of 25(OH)D were elevated by 23 percent in sepsis patients compared to non-septic ones ($p = 0.003$), which indicates a more advanced role of vitamin D status in immunologically related outcomes.

Equally in cardiac patients, the combination of low 25(OH)D and high levels of CRP resulted in a 35 percent increase in SHAP than it would in the absence of the separate influences of each factor (interaction term $p < 0.001$). This fact demonstrates the importance of clinical consideration of both nutritional and inflammatory status simultaneously.

The trends presented by the SHAP values were significant when it came to the predictive ability of vitamin D, as related to time aspects of the clinical scenarios. In the subset of patients in the ICU, we showed that the SHAP variable 25(OH)D was a positive increase in patients with sepsis (23 percent increase, $p = 0.003$) indicating a greater impact of vitamin D status on immunologically related outcomes (Figure 11).

Equally in cardiac patients, the combination of low 25(OH)D and high levels of CRP resulted in a 35 percent increase in SHAP than it would in the absence of the separate influences of each factor (interaction term $p < 0.001$). This fact underscores the need for clinical judgment of the nutritional and inflammatory status (Figure 12).

Population-specific trends of variation in the predictive importance of vitamin D were revealed by demographic stratification of SHAP values. Older patients (>65 years) were also found to have 28% higher SHAP values of low 25(OH)D than younger adults ($p < 0.001$) in accordance with the known age-related tendencies of increased metabolic significance of vitamin D.

The ethnicity analyses showed a difference of 31 and 24 percent in SHAP values of Black and Hispanic patients compared to White patients, respectively (explaining the SHAP difference according to cutaneous production of vitamin D; $p < 0.01$).

Demographic stratification revealed age and ethnic differences in predictive importance of vitamin D as seen in Figure 13. SHAP values were higher in older adults, Black patients and Hispanic patients compared to their comparison groups.

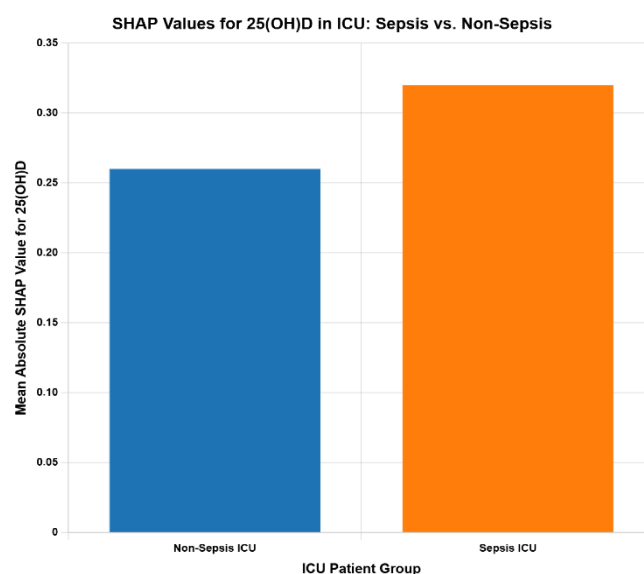


Figure 11. SHAP values for 25(OH)D in ICU: sepsis vs. non-sepsis.

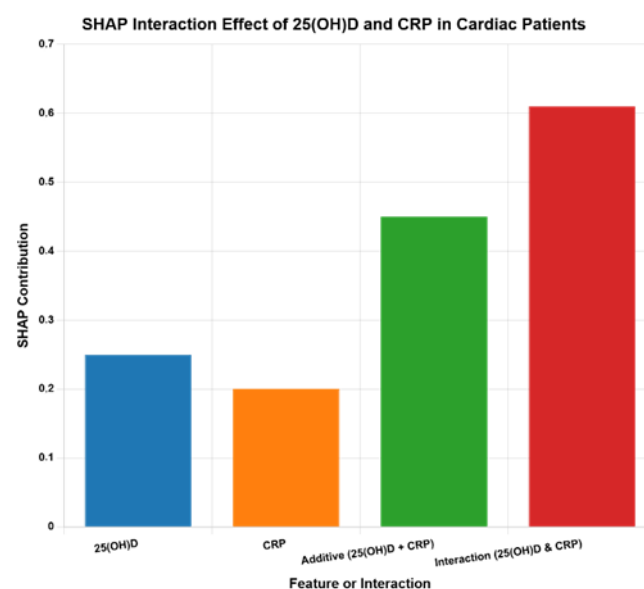


Figure 12. SHAP interaction effect of 25(OH)D and CRP in cardiac patients.

SHAP analysis enhances and aids clinical translation by transforming model results into patient-specific risk-contribution scores that may be incorporated into EHR systems. This way, clinicians can not only see the overall risk for a patient but also the particular role of vitamin D in comparison to other variables; recalculation will be performed dynamically in cases of changes in laboratory information.

In all models, vitamin D is a predictor of importance, which gives it model-neutral support in its clinical risk factor. Linear associations in the Logistic Regression correspond to dose–response and threshold effects, whereas non-linear interactions are represented with other models, which characterize the complexity of biological control.

SHAP can be used to optimize the screening levels of 25(OH)D, objectively measure risk reduction following supplementation, and at the population level to identify high-impact subgroups. This makes SHAP a viable compromise between predictive and precision prevention approaches to the risk of developing a disease related to vitamin D.

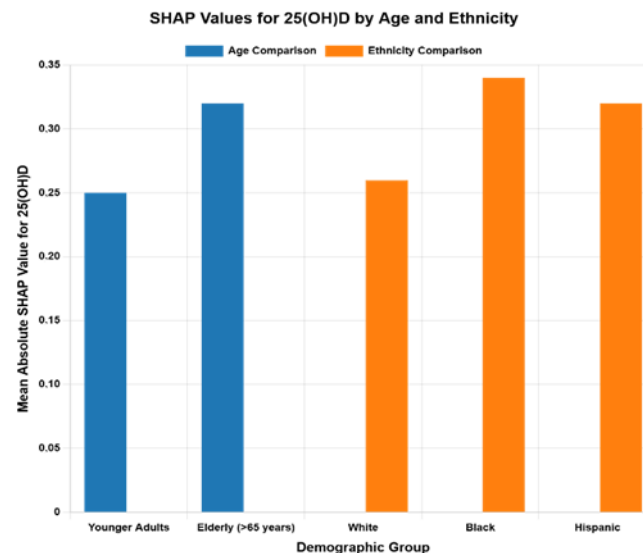


Figure 13. SHAP values for 25(OH)D by age and ethnicity.

5. Discussion

By emphasizing the likely potential of machine learning to assess vitamin D as a predictive biomarker of chronic disease, this study emphasizes the significance of model selection in precision medicine. The excellent performance of the Logistic Regression (LR) disputes the widespread belief that biomedical data requires sophisticated algorithms. In a univariate context, LR leverages underlying threshold-like or dose–response correlations (e.g., deficiency < 20 ng/mL), which agrees with Occam’s Razor. In comparison, Random Forest uses the interaction between features and therefore cannot provide much benefit when there is only one biomarker. SVM and KNN showed a sensitivity/specificity trade-off, which indicates that the model of choice must be clinical-goal-oriented, LR to be used in a wide-screen screening, and more conservative models need to be used in high-specificity triage settings.

Notably, LR has inbuilt interpretability with odds ratios that can easily be estimated, making it easier to win clinician confidence and translatability. This is the opposite of the miscellaneous decision boundary of SVM or the complexity of an ensemble of RF. A middle way between population-level LR screening and targeted non-linear models can be a way to go.

Biologically, the results support the pleiotropic functions of vitamin D in immunomodulation, metabolic functions, and cardiovascular pathways. The resultant threshold sensitivity is consistent with the supplementation studies that have shown value in deficient populations. However, moderate predictive accuracy is also a warning against overinterpretation since vitamin D can be an indicator of more general health conditions and not an entirely causal element. The difference requires advanced causal techniques, such as Mendelian randomization, to explain the distinction.

Unique merits and demerits existed in each of the ML algorithms used. Logistic regression achieved the best results in terms of model interpretability and performance, which made it ideal for clinical use. While both Random Forest and XGBoost were successful in modeling complex interactions between features through their non-linear approach, the downside is that their model interpretability is relatively lower, while SVM performed excellently when dealing with high-dimensional feature vectors.

The result achieved in this research is equivalent to, or in some instances, better than those obtained from previous predictive analysis involving Vitamin D. Previous studies have shown accuracies within the range of 80% to 85% using Random Forest models in the

prediction of vitamin D deficiency and cardiovascular risks [40,41], while deep learning models have been able to achieve an accuracy within the range of 88% to 92% in predicting diseases based on their biomarkers [42,43]. The results of this experiment show Logistic Regression achieving an accuracy of 85.2% and a ROC-AUC of 92.5%; SVM had a ROC-AUC of 90.9%.

The methodology is limited by its univariate design, failure to adjust for confounders, and reliance on single-center data, which limits extrapolation to other populations with different baseline vitamin D levels and assay techniques. External validation and stratified disease subtype analysis are necessary.

All in all, though LR shows good translational possibilities, further studies must include more biomarkers and prospective validation in order to capture the multifactoriality of chronic disease. A combination of interpretability, multimodal modeling, and causal inference will be important to fully realize the vitamin D role in precision prevention.

6. Conclusions

This study recognizes serum 25-hydroxyvitamin D [25(OH)D] as a valuable biomarker of susceptibility to chronic diseases and demonstrates that interpretable machine learning can transform nutritional biology into useful predictions. With the MIMIC-III, the predictive performance of Logistic Regression was superior compared to more complex models, highlighting an important point: in biomarker medicine, biological coherence and interpretability can be of greater importance than the complexity of the algorithms used.

The strong level of dose–response, as well as the considerably increased risk with severe deficiency, highlights the systemic role of vitamin D in immune and metabolic regulation. The findings also indicate that we should not study chronic diseases using a single marker only, since they are caused by a number of things.

The paper challenges the idea that increasingly sophisticated AI is inherently better and advocates the approach that encourages context-sensitive and clinically sensitive model selection via strict validation and open modeling.

The main scientific contribution made by this study can be summarized as the creation and comparative analysis of a variety of machine learning models for chronic disease risk prediction based on serum vitamin D measurements collected from the MIMIC-III dataset. Contrary to many prior works, which have only explored the associations between various factors, the current study examined the use of machine learning for predicting outcomes based on the measurement, demonstrating that interpretable algorithms can perform at an adequate level.

The results of our study prove that vitamin D is not only a risk biomarker but also can be used as a scalable starting point to AI-enhanced precision prevention in chronic disease management around the world.

Future research must focus on expanding the current study with the addition of further nutritional, genetic, laboratory, and lifestyle-related factors to reflect the multi-factorial nature of chronic diseases. It is equally important to validate the models externally with the use of an independent dataset and different patient populations to gauge the models' generalizability. Finally, the application of explainable AI techniques with multimodal clinical data could enhance the prediction performance and utility of the models.

Author Contributions: Writing—original draft preparation S.C.; Validation-Supervision M.T.; Supervision H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by the Moroccan School of Engineering Sciences EMSI Casablanca.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MIMIC-III	Medical Information Mart for Intensive Care III
ML	Machine Learning
LR	Logistic Regression
SVM	Support Vector Machines
RF	Random Forest
KNN	K-Nearest Neighbors
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
AI	Artificial Intelligence
NCD	Non-Communicable Diseases
WHO	World Health Organization
UVB	Ultraviolet-B radiation
EHR	Electronic Health Records
VDR	Vitamin D Receptor
RAAS	Renin–Angiotensin–Aldosterone System
CCS	Clinical Classifications Software
ICD-9	International Classification of Diseases, Ninth Revision
SIRS	Systemic Inflammatory Response Syndrome
RFE	Recursive Feature Elimination

References

1. World Health Organization (WHO). *Noncommunicable Diseases*; WHO Fact Sheet: Geneva, Switzerland, 2023. Available online: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> (accessed on 7 May 2026).
2. Chebbawi, S.; Tabaa, M.; Badir, H. Investigating the impact of Vitamin D on chronic diseases using Machine Learning. *Procedia Comput. Sci.* **2025**, *257*, 374–381. [\[CrossRef\]](#)
3. Amrein, K.; Scherkl, M.; Hoffmann, M.; Neuwersch-Sommeregger, S.; Köstenberger, M.; Berisha, A.T.; Martucci, G.; Pilz, S.; Malle, O.; Dobnig, H.; et al. Vitamin D Deficiency 2.0: An Update on the Current Status Worldwide. *Eur. J. Clin. Nutr.* **2020**, *74*, 1498–1513. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Cui, A.; Zhang, T.; Xiao, P.; Fan, Z.; Wang, H.; Zhuang, Y. Global and Regional Prevalence of Vitamin D Deficiency in Population-Based Studies from 2000 to 2022: A Pooled Analysis of 7.9 Million Participants. *Front. Nutr.* **2023**, *10*, 1070808. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Holick, M.F. Vitamin D Deficiency. *N. Engl. J. Med.* **2007**, *357*, 266–281. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Nair, R.; Maseeh, A. Vitamin D: The “sunshine” vitamin. *J. Pharmacol. Pharmacother.* **2012**, *3*, 118–126. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Bouillon, R.; Marcocci, C.; Carmeliet, G.; Bikle, D.; White, J.H.; Dawson-Hughes, B.; Lips, P.; Munns, C.F.; Lazaretti-Castro, M.; Giustina, A.; et al. The Health Effects of Vitamin D Supplementation: Evidence from Human Studies. *Nat. Rev. Endocrinol.* **2021**, *17*, 37–50. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Charoenngam, N.; Holick, M.F. Immunologic Effects of Vitamin D on Human Health and Disease. *Nutrients* **2020**, *12*, 2097. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Mailhot, G.; White, J.H. Vitamin D and Immunity in Infants and Children. *Nutrients* **2020**, *12*, 1233. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Theodoratou, E.; Tzoulaki, I.; Zgaga, L.; Ioannidis, J.P.A. Vitamin D and Multiple Health Outcomes: Umbrella Review of Systematic Reviews and Meta-Analyses of Observational Studies and Randomised Trials. *BMJ* **2014**, *348*, g2035. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Garland, F.C.; Garland, C.F.; Gorham, E.D.; Young, J.F. Geographic Variation in Breast Cancer Mortality in the United States: A Hypothesis Involving Exposure to Solar Radiation. *Prev. Med.* **1980**, *9*, 614–622. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Wang, T.J.; Pencina, M.J.; Booth, S.L.; Jacques, P.F.; Ingelsson, E.; Lanier, K.; Benjamin, E.J.; D’Agostino, R.B.; Wolf, M.; Vasan, R.S. Vitamin D Deficiency and Risk of Cardiovascular Disease. *Circulation* **2008**, *117*, 503–511. [\[CrossRef\]](#) [\[PubMed\]](#)

13. Chen, S.; Sun, Y.; Agrawal, D.K. Vitamin D Deficiency and Essential Hypertension. *J. Am. Soc. Hypertens.* **2015**, *9*, 885–901. [[CrossRef](#)] [[PubMed](#)]
14. Pilz, S.; Kienreich, K.; Tomaschitz, A.; Lerchbaum, E.; Meinitzer, A.; März, W.; Zittermann, A.; Dekker, J.M. Vitamin D and Cardiovascular Disease: Update and Outlook. *Scand. J. Clin. Lab. Invest. Suppl.* **2012**, *72*, 83–91. [[CrossRef](#)] [[PubMed](#)]
15. Pittas, A.G.; Dawson-Hughes, B.; Sheehan, P.; Ware, J.H.; Knowler, W.C.; Aroda, V.R.; Brodsky, I.; Ceglia, L.; Chadha, C.; Chatterjee, R.; et al. Vitamin D Supplementation and Prevention of Type 2 Diabetes. *N. Engl. J. Med.* **2019**, *381*, 520–530. [[CrossRef](#)] [[PubMed](#)]
16. Pramono, A.; Jocken, J.W.E.; Blaak, E.E.; van Baak, M.A. The Effect of Vitamin D Supplementation on Insulin Sensitivity: A Systematic Review and Meta-analysis. *Diabetes Care* **2020**, *43*, 1659–1669. [[CrossRef](#)] [[PubMed](#)]
17. Keum, N.; Lee, D.H.; Greenwood, D.C.; Manson, J.E.; Giovannucci, E. Vitamin D supplementation and total cancer incidence and mortality: A meta-analysis of randomized controlled trials. *Ann. Oncol.* **2019**, *30*, 733–743. [[CrossRef](#)] [[PubMed](#)]
18. Feldman, D.; Krishnan, A.V.; Swami, S.; Giovannucci, E.; Feldman, B.J. The Role of Vitamin D in Reducing Cancer Risk and Progression. *Nat. Rev. Cancer* **2014**, *14*, 342–357. [[CrossRef](#)] [[PubMed](#)]
19. Wang, H.; Chen, W.; Li, D.; Yin, X.; Zhang, X.; Olsen, N.; Zheng, S.G. Vitamin D and Chronic Diseases. *Aging Dis.* **2017**, *8*, 346–353. [[CrossRef](#)] [[PubMed](#)]
20. Hahn, J.; Cook, N.R.; Alexander, E.K.; Friedman, S.; Walter, J.; Bubes, V.; Kotler, G.; Lee, I.-M.; Manson, J.E.; Costenbader, K.H. Vitamin D and marine omega 3 fatty acid supplementation and incident autoimmune disease: VITAL randomized controlled trial. *BMJ* **2022**, *376*, e066452. [[CrossRef](#)] [[PubMed](#)]
21. Grant, W.B.; Wimalawansa, S.J.; Pludowski, P.; Cheng, R.Z. Vitamin D: Evidence-Based Health Benefits and Recommendations for Population Guidelines. *Nutrients* **2025**, *17*, 277. [[CrossRef](#)] [[PubMed](#)]
22. Roth, D.E.; Morris, S.K.; Zlotkin, S.; Gernand, A.D.; Ahmed, T.; Shanta, S.S.; Papp, E.; Korsiak, J.; Shi, J.; Islam, M.M.; et al. Vitamin D Supplementation in Pregnancy and Lactation and Infant Growth. *N. Engl. J. Med.* **2018**, *379*, 535–546. [[CrossRef](#)] [[PubMed](#)]
23. Wimalawansa, S.J. Vitamin D Deficiency: Effects on Oxidative Stress, Epigenetics, Gene Regulation, and Aging. *Biology* **2019**, *8*, 30. [[CrossRef](#)] [[PubMed](#)]
24. Bouillon, R. Extra-Skeletal Effects of Vitamin D. *Front. Horm. Res.* **2018**, *50*, 72–88. [[CrossRef](#)] [[PubMed](#)]
25. Waterhouse, M.; Hope, B.; Krause, L.; Morrison, M.; Protani, M.M.; Zakrzewski, M.; Neale, R.E. Vitamin D and the Gut Microbiome: A Systematic Review of In Vivo Studies. *Eur. J. Nutr.* **2019**, *58*, 2895–2910. [[CrossRef](#)] [[PubMed](#)]
26. Autier, P.; Boniol, M.; Pizot, C.; Mullie, P. Vitamin D Status and Ill Health: A Systematic Review. *Lancet Diabetes Endocrinol.* **2014**, *2*, 76–89. [[CrossRef](#)] [[PubMed](#)]
27. Manson, J.E.; Cook, N.R.; Lee, I.-M.; Christen, W.; Bassuk, S.S.; Mora, S.; Gibson, H.; Gordon, D.; Copeland, T.; D’Agostino, D.; et al. Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. *N. Engl. J. Med.* **2019**, *380*, 33–44. [[CrossRef](#)] [[PubMed](#)]
28. Wacker, M.; Holick, M.F. Vitamin D—Effects on Skeletal and Extraskelatal Health and the Need for Supplementation. *Nutrients* **2013**, *5*, 111–148. [[CrossRef](#)] [[PubMed](#)]
29. Zmijewski, M.A. Vitamin D and Human Health. *Int. J. Mol. Sci.* **2019**, *20*, 145. [[CrossRef](#)] [[PubMed](#)]
30. Bikle, D.D. Vitamin D: Newer Concepts of Its Metabolism and Function at the Basic and Clinical Level. *Endocr. Rev.* **2020**, *41*, 345–366. [[CrossRef](#)] [[PubMed](#)]
31. Zhao, S.; Gardner, M.D.; Engmann, J.; Anderson, D.; Michos, E.D. Vitamin D and Major Chronic Diseases. *Trends Endocrinol. Metab.* **2024**, *35*, 1050–1061. [[CrossRef](#)] [[PubMed](#)]
32. Carlberg, C.; Vukovic, A.V. Vitamin D in the Context of Evolution. *Nutrients* **2022**, *14*, 3018. [[CrossRef](#)] [[PubMed](#)]
33. Subramanian, M.; Wojtuszczy, A.; Favre, L.; Boughorbel, S.; Shan, J.; Letaief, K.B.; Pitteloud, N.; Chouchane, L. Precision Medicine in the Era of Artificial Intelligence: Implications in Chronic Disease Management. *J. Transl. Med.* **2020**, *18*, 472. [[CrossRef](#)] [[PubMed](#)]
34. Sosa-Holwerda, A.; Park, O.-H.; Albracht-Schulte, K.; Niraula, S.; Thompson, L.; Oldewage-Theron, W. The Role of Artificial Intelligence in Nutrition Research: A Scoping Review. *Nutrients* **2024**, *16*, 2066. [[CrossRef](#)] [[PubMed](#)]
35. Panayotova, G.G. Artificial Intelligence in Nutrition and Dietetics: A Comprehensive Review of Current Research. *Healthcare* **2025**, *13*, 2579. [[CrossRef](#)] [[PubMed](#)]
36. Eşsiz, U.E.; Yüregir, O.H.; Saraç, E. Applying Data Mining Techniques to Predict Vitamin D Deficiency in Diabetic Patients. *Health Inform. J.* **2023**, *29*. [[CrossRef](#)] [[PubMed](#)]
37. Pang, J.; Yang, C.; Liu, J.; Wang, Z.; Tao, X.; Cao, Z. Correlation between Vitamin D Metabolic Pathway-Related Gene Polymorphisms and Cardiovascular Disease. *Food Funct.* **2024**, *15*, 11342. [[CrossRef](#)] [[PubMed](#)]
38. Sancar, N.; Tabrizi, S.S. Machine Learning Approach for the Detection of Vitamin D Level: A Comparative Study. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 219. [[CrossRef](#)] [[PubMed](#)]
39. Wang, C.-K.; Chang, C.-Y.; Chu, T.-W.; Liang, Y.-J. Using Machine Learning to Identify the Relationships between Vitamin D and Demographic, Biochemical, and Lifestyle Factors. *Life* **2023**, *13*, 2257. [[CrossRef](#)] [[PubMed](#)]

40. Patino-Alonso, C.; Gómez-Sánchez, M.; Gómez-Sánchez, L.; Sánchez Salgado, B.; Rodríguez-Sánchez, E.; García-Ortiz, L.; Gómez-Marcos, M.A. Predictive Ability of Machine-Learning Methods for Vitamin D Deficiency Prediction by Anthropometric Parameters. *Mathematics* **2022**, *10*, 616. [[CrossRef](#)]
41. Guo, J.; He, Q.; Li, Y. Machine Learning-Based Prediction of Vitamin D Deficiency. *Front. Endocrinol.* **2024**, *15*, 1327058. [[CrossRef](#)] [[PubMed](#)]
42. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [[CrossRef](#)] [[PubMed](#)]
43. Alzahrani, A.; Asghar, M.Z. Enhancing the Prediction of Vitamin D Deficiency Levels Using an Integrated Approach of Deep Learning and Evolutionary Computing. *PeerJ Comput. Sci.* **2025**, *11*, e2698. [[CrossRef](#)] [[PubMed](#)]
44. Jiménez-Gaona, Y.; Vivanco-Galván, O.; Castillo-Malla, D.; Vivanco-Gualán, I.; Díaz-Guzmán, P. VITA-D: A Radiomic Web Tool for Predicting Vitamin D Deficiency Levels. *Appl. Sci.* **2025**, *15*, 1798. [[CrossRef](#)]
45. Chauhan, D.; Yadav, A.K.; Bhatia, D.; Solanki, P.R. Global Impact of Vitamin D Deficiency and Innovative Biosensing Technologies. *Chem. Eng. J.* **2025**, *506*, 159790. [[CrossRef](#)]
46. Wu, X.; Oniani, D.; Shao, Z.; Arciero, P.; Sivarajkumar, S.; Hilsman, J.; Mohr, A.E.; Ibe, S.; Moharir, M.; Li, L.-J.; et al. A Scoping Review of Artificial Intelligence for Precision Nutrition. *Adv. Nutr.* **2025**, *16*, 100398. [[CrossRef](#)] [[PubMed](#)]
47. Mavar, M.; Sorić, T.; Bagarić, E.; Sarić, A.; Matek Sarić, M. The Power of Vitamin D: Is the Future in Precision Nutrition through Personalized Supplementation Plans? *Nutrients* **2024**, *16*, 1176. [[CrossRef](#)] [[PubMed](#)]
48. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and Accurate Deep Learning with Electronic Health Records. *npj Digit. Med.* **2018**, *1*, 18. [[CrossRef](#)] [[PubMed](#)]
49. Giustina, A.; Bilezikian, J.P.; Adler, R.A.; Banfi, G.; Bikle, D.D.; Binkley, N.C.; Bollerslev, J.; Bouillon, R.; Brandi, M.L.; Casanueva, F.F.; et al. Consensus Statement on Vitamin D Status Assessment and Supplementation: Whys, Whens, and Hows. *Endocr. Rev.* **2024**, *45*, 625–654. [[CrossRef](#)] [[PubMed](#)]
50. Sargiotis, G.C. Applications of Artificial Intelligence in Clinical Studies: Ethical Dilemmas. *Bioethica* **2023**, *9*, 123–130.
51. Pan, M.; Li, R.; Wei, J.; Peng, H.; Hu, Z.; Xiong, Y.; Li, N.; Guo, Y.; Gu, W.; Liu, H. Application of Artificial Intelligence in the Health Management of Chronic Disease: Bibliometric Analysis. *Front. Med.* **2025**, *11*, 1506641. [[CrossRef](#)] [[PubMed](#)]
52. Quan, W.; Zhou, J.; Wang, J.; Huang, J.; Du, L. Machine Learning-Driven Precision Nutrition: A Paradigm Evolution in Dietary Assessment and Intervention. *Nutrients* **2026**, *18*, 45. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.