International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 8, Issue 4
DOI: https://doi.org/10.15157/IJITIS.2025.8.4.911-936
Received: 23.09.2025; Revised: 30.10.2025; Accepted: 17.11.2025

Research Article

Predicting Vitamin D Levels Using Ordinal Logistic Regression, Gaussian Process Regression and ARIMA: A Comparative Study

Edlira Lashi[1*], Klea Lashi[2], Hasanien K Kuba[3], Andres Annuk[4*], Ambrozia Itellari[5], Hussein Alkattan[6,7], Mostafa Abotaleb[7]

[1] Faculty of Medical Science, Albanian University, Tirana, Albania
[2] Department of Clinical Biochemistry, Mother Teresa University Hospital Center, Tirana, Albania
[3] College of Biomedical Informatics, University of Information Technology and Communications (UoITC), Baghdad, Iraq
[4] Institute of Forestry and Engineering, Estonian University of Life Sciences, Tartu, Estonia
[5] School of Medicine, Lake Erie College of Osteopathic Medicine, Florida, United States of America
[6] Department of System Programming, South Ural State University, Chelyabinsk, Russia
[7] Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq,
[8] Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, Russia
***e.lashi@albanianuniversity.edu.al; andres.annuk@emu.ee**

Abstract

Vitamin D deficiency is a common health condition that increases the risk of metabolic, cardiovascular, and musculoskeletal disorders. Many individuals are unaware of their vitamin D deficiency. In this work, we develop and present three complementary machine learning models to explore Vitamin D levels based on regular healthcare data. The dataset consists of anonymized patient records with demographic features, clinical indicators, and laboratory measurements of serum 25(OH)D. It is taken from a healthcare setting and pre-processed to eliminate absent or inconsistent results. Vitamin D level variables were transformed into ordered, clinical categories: severe deficiency, deficiency, insufficiency, and sufficiency. However, for regression and time-series forecasting, the original continuous concentration, measured in ng/mL, was preserved together with monthly averages. A proportional odds Ordinal Logistic Regression model was used to figure out Vitamin D status. The best overall performance was an accuracy of 0.77, a macro recall of 0.76, and an F2-score of 0.78. Most of the mistakes were made between categories that were next to each other. We utilized Gaussian Process Regression to predict continuous Vitamin D concentration. The results were $R^2 = 0.79$, MAE = 2.3 ng/mL, and RMSE = 3.4 ng/mL, which means that the model can get close to laboratory values with clinically acceptable error. To capture temporal dynamics, an ARIMA model was fitted to monthly mean Vitamin D levels and showed the best performance with $R^2 = 0.82$, MAE = 2.0 ng/mL and RMSE = 3.1 ng/mL, accurately recreating the observed seasonal pattern.

Keywords: Vitamin D; Healthcare Data; Machine Learning; Ordinal Logistic Regression; Gaussian Process Regression; ARIMA; Time-series Forecasting; Clinical Decision Support

INTRODUCTION

Vitamin D is increasingly realized as an important vitamin in human health, besides its long-established role of mineralization of the bone, due to its long-lasting influence on immunity, cardiovascular integrity, endocrine control, and prevention of chronic disease.

Adequate amounts of this fat-soluble vitamin are required for calcium and phosphorus metabolism, skeletal function, and muscular function; conversely, its deficiency has been unequivocally associated with numerous clinical disorders [1-6]. A number of studies emphasize the need to identify the determinants of Vitamin D level and develop accurate prediction models to guide treatment. Traditionally, Vitamin D is obtained from diet or endogenously synthesized after exposure to UVB radiation [7-11]. Holick [3] emphasized that the sun is the primary natural source, yet the balance between proper exposure and risk of skin cancer is a significant public health concern.

Lifestyle changes in modern times, indoor dwelling, and the use of sunscreens have resulted in the global deficiency of vitamin D. Recent epidemiological studies have shown the burden of illness, and a meta-analysis of 7.9 million patients from 2000 to 2022 has found that Vitamin D deficiency is highly prevalent [12].

The physiological function of Vitamin D is not restricted to bone metabolism.

Calcitriol, the bioactive metabolite, affects muscle metabolism and protein synthesis, thereby increasing muscular strength and reducing the risk of sarcopenia. Popa et al. explained the complex relationship between Vitamin D deficiency, obesity, and inflammation, and suggested that deficiency may promote proinflammatory pathways. Vitamin D is an important regulator of both innate and adaptive immunity, and its deficiency has been linked to autoimmune diseases and reduced resistance to infections. Parkhe et al. [2] explained the immunomodulatory role of vitamins, emphasizing their ability to enhance host resistance to diseases and respond to emerging infectious challenges. Vitamin D deficiency presents with many clinical manifestations.

A correlation between Vitamin D deficiency and higher HbA1C in people with type 2 diabetes mellitus was identified by Zhao et al. [8]. In them, there is an association with metabolic disease.

Ingles et al. [6] explained how supplemental vitamins, including Vitamin D, could influence cardiovascular risk factors and outcomes in cardiology. Alagacone et al. [7] showed that Vitamin D deficiency is related to resistant hypertension, proving the systemic effects of the vitamin.

These three studies taken together suggest that Vitamin D is an important determinant of health and disease burden. Population-level prevention has emphasized food fortification. Niedermaier et al. [9] estimated that effective Vitamin D fortification programs in European countries could substantially reduce cancer mortality, highlighting the unexploited potential of dietary intervention. The correct dosage, form, and route of administration must yet be elucidated, since Bilezikian et al. [10] accounted for the impossibility of employing a single approach to all groups and therapeutic conditions. The metabolic complexity of Vitamin D is also supported by the fact that there are epimers, as seen from Al-Zohily et al. [5], who asserted that these compounds interfere with laboratory measurement and interpretation.

Authors in [13-15] expressed concern on the undertreatment and underdiagnosis of Vitamin D deficiency in bone metabolism and osteoporosis and suggest more robust diagnostic and therapeutic strategies. Meanwhile, other writers associated Vitamin D deficiency with the activity and inability of rheumatoid arthritis [16-18] and the clinical importance of decreased levels of Vitamin D [19]. Bechrouri et al. [19] performed a comparative investigation of statistical models in the estimation of Vitamin D levels, highlighting the importance of quantification in clinical decision-making. Karamizadeh et al. [14] established that serum levels of 25-hydroxyvitamin D can be accurately estimated using linear regression and machine learning algorithms, with better results compared to conventional approaches. Machine learning platforms have the capability for multiple factors like demographic, biochemical, and lifestyle information to be integrated to predict the risk of deficiency and implement appropriate tailored interventions. The future of predictive medicine is through assembling machine models that are combining the strengths of multiple algorithms. This approach diminishes the bias and variability, reduces error rates, and generalizes better to populations.

Machine learning approaches to Vitamin D research prove to be significantly effective in predicting serum levels and classifying individuals into clinically relevant categories of deficiency, inadequacy, and adequacy. Predictive results are necessary to define dietary recommendations, build fortification strategy policies, and plan therapeutic treatment for populations at risk. Vitamin D is a valuable biomarker at the interface of nutrition, endocrinology, immunology, and prevention of chronic disease. New evidence explains its physiological function and the chronic global problem of deficiency. Machine learning and hybrid predictive models offer new solutions to these challenges through facilitating precise, evidence-based monitoring and risk classification. This study establishes a hybrid machine learning model for the prediction of Vitamin D levels based on demographic and biochemical data, with an aim to improve the diagnostic accuracy and therapeutic decision-making, in accordance with earlier work [15–19].

RELATED WORK

Recent scientific advancements in Vitamin D have greatly enhanced its clinical and physiological significance, computational and machine learning application in detection and prediction.

Its toxicological significance, psychological effects, cardiometabolic outcomes, and predictive modelling of the Vitamin D level have been studied by many studies. Concurrently, the scientific method increasingly employs ensemble learning and hybrid methods in disease prediction to establish the amenability of contemporary algorithms to medical use.

Toxicology studies discover that Vitamin D is required for health but over-supplementation is highly dangerous, particularly in children.

Levita et al. [19] performed rigorous research and case reports on Vitamin D toxicity in children in favor of cautious dosing within clinical contexts. aside from toxicity, Vitamin D has also been associated with mental health. Jahan-Mihan et al. [20] showed it to be effective for the prevention and treatment of depression and seasonal affective disorder in adults, an affirmation of earlier work by Casseb et al. [21] that confirmed the effectiveness of Vitamin D in preventing depression and anxiety. It looks at various applications of Vitamin D involving its impact on mental and skeletal health [22-26].

The cardiometabolic role of Vitamin D has been a subject of continuous research. Jääskeläinen et al. [27] investigated Vitamin D status as a predictor of weight gain or waist circumference increase in the Finnish prospective population. Strong connections among obesity metrics were observed. Davies et al. 30 presented compelling evidence for the causal relationship between Vitamin D level and all COVID-19 outcomes. Significant immunomodulatory effects were observed. These data clearly demonstrate that Vitamin D insufficiency is both a dietary inadequacy and a condition associated with increased illness and infection risk. As machine learning applications continue to develop in healthcare, computational techniques are used for Vitamin D prediction. Sancar and Tabrizi 28 conducted a comparative analysis of ensemble-based and machine learning models for the prediction of Vitamin D and emphasized the need for support of the ensemble-based techniques.

Guo et al. [29] utilized support vector regression (SVR) to forecast Vitamin D level in the Ausimmune Study cohort, highlighting the potential of non-linear regression models. Islam et al. [31] developed an interpretability-centric ensemble method for diagnosing Vitamin D deficiency, valuing prediction model accuracy and interpretability over all else. Such machine learning innovations have the potential to enhance classical clinical testing through non-invasive, evidence-based prediction. Vitamin D use experiments with sophisticated computer simulation also reference other chronic disease prediction studies. Ensemble and hybrid learning methods are commonly used for cardiovascular and metabolic disorder prediction. Ensemble boosting model was suggested by Ganie et al. [22] for cardiac disease prediction with significant improvement from conventional techniques. Noor et al. [23] extended previous work in the development of a stacking model that combines balancing methods and dimensionality reduction of feature space, thereby efficiently minimizing imbalances in cardiac data sets. Mondal et al. [24] extended this by developing a two-stage stacked machine learning model to estimate the risk of heart disease and showed enhanced efficiency in clinical practices. Ensemble learning has been used in oncology and chronic disease management. Jadoon et al. [25] have developed a deep learning ensemble classifier for multi-modal breast cancer prediction, enhancing diagnostic accuracy through data modality integration. Al-Jamimi [26] proposed an ensemble learning and feature engineering approach to chronic disease prediction, effectively integrating data preprocessing with ensembles of classifiers. These researches provide a methodological answer for how to enhance ensemble methods in Vitamin D prediction using the same heterogeneous clinical and demographic data to maximize precision. The implications of these computing efforts are extensive.

Predictive modelling of Vitamin D status will facilitate early detection of insufficiency, direct supplementation regimens, and enhance risk management for over-supplementation. Levita et al.'s [19] toxicity study highlights the prevention of pediatric overdose, and this can be most effectively done through the use of computer algorithms that provide personalized dosage recommendations. These models such as Islam et al. [31] and Sancar and Tabrizi [28] may be improved in diagnostic processes through the provision of timely and interpretable Vitamin D predictions, particularly in low-resource settings. Comparative literary analysis of the studies points towards the shift from conventional statistical models to machine learning and ensemble hybrid models. Guo et al. [29] originally showed evidence for SVR; however, more recent studies, such as by Islam et al. [31], propose a new paradigm of ensemble methods where heterogeneous algorithms work together to build resilience. This is in agreement with new developments in heart disease [22–24] and cancer prediction [25], where stacking, boosting, and multi-modal ensembles are dominant methods.

The uniformity of approach for all disciplines means that ensemble machine learning is capable of solving other biomedical prediction issues, such as Vitamin D deficiency. There is described in the literature a bifurcated strategy. Biomedical science is rationally investigating the widespread clinical relevance of Vitamin D in bone and metabolic disease and in psychological and immunological engagement [30]. Conversely, computational science has increasingly become concerned with the application of machine learning techniques to quantify, forecast, and classify Vitamin D status [31].

The objective of this study is to contribute to the literature by building a hybrid machine learning model which brings together comprehension from two dissimilar bodies of knowledge. It brings together regression and classification techniques to forecast continuous vitamin D levels and classify patients into clinical categories, thereby merging the interpretability of statistical models with the accuracy of ensemble learning.

DATA AND METHODOLOGY

Dataset

This study utilized routinely collected healthcare data from adult patients who underwent serum 25-hydroxyvitamin D (25(OH)D) testing in a clinical laboratory. After the application of record inclusion and exclusion criteria, the dataset contained a total of 520 anonymized records, each corresponding to a unique patient visit and including demographic and clinical information related to Vitamin D status.

The target variable, serum 25(OH)D concentration (ng/mL), was determined using typical immunoassay methods in the hospital lab. We further categorized the continuous 25(OH)D levels into four ordered groups using standard clinical thresholds that indicate Vitamin D status:

- Severe deficiency: 25(OH)D < 10 ng/mL
- Not enough: 10–19.9 ng/mL
- Not sufficient: 20–29.9 ng/mL
- Sufficient: ≥ 30 ng/mL

These groups were used as the dependent variable by the ordinal classification model. The input features were selected from healthcare data that is typically available and included:

- Demographic factors such as age in years, sex (male/female), and body mass index are all considered here.
- The clinical variables of interest are: presence of a chronic condition (e.g., diabetes or hypertension); intake of Vitamin D supplementation (yes/no); and some biochemical markers (e.g., calcium or creatinine-if available).
- Temporal/contextual variables included the date of the blood sampling, which was further transformed into month and season - winter, spring, summer, autumn - considering that Vitamin D may have seasonal variations.

All the identifiers were taken out before the analysis; therefore, there was no direct personal information within the dataset that kept patients' privacy.

Data Preprocessing

Different preprocessing steps were taken prior to model building:

- Data cleaning: Records without 25(OH)D values were excluded. Predictor variables with more than 20% missing values were excluded from the analysis. Missing values for continuous variables (for instance, BMI) were imputed using the median of that respective feature, while missing binary variables (for example, supplements yes/no) were imputed with the mode.
- Outlier handling: Improbable 25(OH)D values-for example, < 3 ng/mL or > 120 ng/mL-were considered measurement errors and were excluded. Extreme outliers in the continuous predictors were minimized at the 1st and 99th percentiles to reduce undue influence without affecting the overall distribution.
- Feature encoding and transformation: Categorical factors (gender, chronic illness, supplements, season) were encoded using dummy variables. Age and BMI were retained as continuous variables. In the regression models (GPR and ARIMA), continuous predictors were standardized to mean zero and variance one, and ordinal categories were encoded as numeric values (1-4) in ascending order of Vitamin D sufficiency.
- The dataset was randomly divided on the patient level into training and test subsets in a ratio of 70:30. The training set was used to fit the models and optimize their hyperparameters, while the test set was reserved for the final assessment of their performance to avoid optimistic bias. Construction of time-series for ARIMA: The monthly mean 25(OH)D levels were calculated by aggregating all data for each

calendar month throughout the period. This resulted in a univariate monthly time series of Vitamin D concentration, which was then used to fit and evaluate the ARIMA model.

Figure 1 shows the detailed workflow used to build the Vitamin D prediction framework using healthcare data.

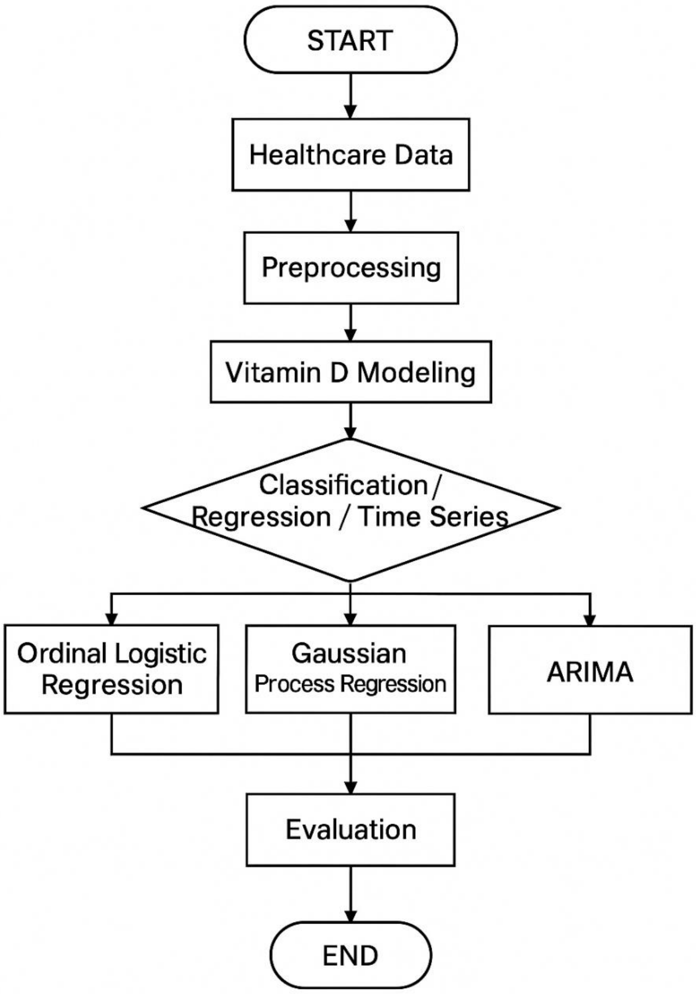

Figure 1. Workflow of Vitamin D Modelling Using Healthcare Data.

The process starts at START, where anonymized healthcare records are collected and summarized under the node Healthcare Data. The raw records advance to the Preprocessing step, which involves handling missing values, examining outliers, encoding variables, and normalizing features to generate a clean analytical dataset. The cleaned dataset moves to the Vitamin D Modelling block, which leads to a decision point called Classification / Regression / Time Series. This indicates that the same data enables three related tasks: categorical classification of Vitamin D status, continuous regression of serum levels, and temporal forecasting of monthly trends. Off this decision node, the process splits into three modelling routes comprising Ordinal Logistic Regression for ordered status classification, Gaussian Process Regression for continuous concentration prediction, and ARIMA for time-series modelling aggregated monthly Vitamin D levels. The results of the three models pass through the Evaluation block, where various performance metrics are calculated and compared.

Ordinal Logistic Regression (Proportional Odds Model)

OLR with a proportional odds link function was used to classify patients into the four ranked Vitamin D status categories. We assumed that the effects of predictors were the same across each of the cumulative logits, thus adhering to the proportionate odds assumption. Predictors included age, sex, body mass index (BMI), chronic disease status, supplements, and seasonality.

Model estimation was done by maximum likelihood. The proportionate odds assumption was checked by routine diagnostics, which did not show significant violations. The fitted OLR model provides cumulative probabilities for every category, which were transformed into the most likely class label for the purpose of performance evaluation.

Let the dataset be

$$\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N} \tag{1}$$

where $\mathrm{x}_i \subset R^p$ is the feature vector for patient i, and $y_i \in \{1,2,\dots,J\}$ is the ordered Vitamin D status (e.g., severe deficiency, deficiency, insufficiency, sufficiency).

The cumulative probability up to category j is:

$$\pi_{ij}^{(<)} = P(Y_i \leq j \mid \mathrm{x}_i), j = 1, \dots, J-1 \tag{2}$$

The corresponding cumulative odds are:

$$\mathrm{Odds}(Y_i \leq j \mid \mathrm{x}_i) = \frac{\pi_{ij}^{(S)}}{1-\pi_{ij}^{(S)}} \tag{3}$$

The proportional odds model assumes a linear predictor in the log-odds scale:

$$\log\left(\frac{\pi_{ij}^{(S)}}{1-\pi_{ij}^{(S)}}\right) = \alpha_j - \mathrm{x}_i^{\top}\beta, j = 1, \dots, J-1 \tag{4}$$

where α_j are category-specific intercepts (cut-points) and $\beta \in \mathrm{R}^p$ is the common slope vector.

Rearranging (4), the cumulative probability can be written as:

$$\pi_{ij}^{(\leq)} = P(Y_i \leq j \mid \mathrm{x}_i) = \frac{\exp(\alpha_j - \mathrm{x}_i^{\top}\beta)}{1+\exp(\alpha_j - \mathrm{x}_i^{\top}\beta)} \tag{5}$$

The category probability for the exact class j is the difference of cumulative probabilities:

$$\pi_{ij} = P(Y_i = j \mid \mathrm{x}_i) = \pi_{ij}^{(S)} - \pi_{i,j-1}^{(S)}, j = 2, \dots, J-1 \tag{6}$$

with the boundary cases

$$\pi_{i1} = \pi_{i1}^{(\zeta)}, \pi_{i,J} = 1 - \pi_{i,J-1}^{(\zeta)} \tag{7}$$

An important interpretation is the odds ratio for a one-unit change in predictor x_k:

$$\mathrm{OR}_k = \exp(-\beta_k) \tag{8}$$

which is assumed to be constant across all cumulative logits (proportional odds assumption).

The likelihood contribution for observation i is:

$$L_i(\theta) = \prod_{j=1}^{J} \pi_{ij}^{I(y-j)} \tag{9}$$

where $\theta = (\alpha_1, \dots, \alpha_{J-1}, \beta)$ and $\mathbb{I}(\cdot)$ is the indicator function.

The log-likelihood over all observations is:

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} \mathbb{I}(y_i = j) \log \pi_{ij} \tag{10}$$

The score vector (gradient of the log-likelihood) is

$$\mathrm{U}(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{N} \mathrm{s}_i(\theta) \tag{11}$$

where $\mathrm{s}_i(\theta)$ collects the partial derivatives w.r.t. α_j and β. Parameter estimates are obtained by solving.

$$\mathrm{U}(\theta) = 0 \tag{12}$$

using a numerical routine, e.g. Newton-Raphson or iteratively reweighted least squares. Given a new patient with features x_*, the predicted category is:

$$\hat{y}_* = \arg \max_{j \in \{1, \dots, J\}} \hat{\pi}_{*j} \tag{13}$$

where $\hat{\pi}_{*j}$ are obtained from (5)-(7) using the estimated parameters. Optionally, one can define an expected ordinal score:

$$\mathbb{E}[Y_* \mid \mathrm{x}_*] = \sum_{j=1}^{J} j \hat{\pi}_{*j} \tag{14}$$

which provides a continuous severity index of Vitamin D deficiency.

Gaussian Process Regression

GPR was applied to the same set of predictors to model the continuous concentration of 25(OH)D. GPR views the underlying regression function as a sample from a Gaussian process, which is specified by a mean function and a covariance (kernel) function. The present study used a zero mean function and a squared exponential kernel with automatic relevance determination.

The kernel's hyperparameters-length scales and noise variance-were determined by maximizing the marginal likelihood on the training dataset. GPR was chosen because of its flexibility, non-parametric nature, and because it can return point forecasts and uncertainty estimates-an important requirement for clinical decisions.

For continuous Vitamin D concentration, we assume the regression model:

$$y_i = f(\mathrm{x}_i) + \varepsilon_i \tag{15}$$

where $f(\cdot)$ is an unknown function and $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ is i.i.d. Gaussian noise.

A Gaussian Process prior is placed on:

$$f(\mathrm{x}) \sim \mathcal{GP}\big(m(\mathrm{x}), k(\mathrm{x}, \mathrm{x}')\big) \tag{16}$$

where $m(\mathrm{x})$ is the mean function (often set to zero) and $k(\mathrm{x}, \mathrm{x}')$ is a positive definite covariance (kernel) function.

In this work we may use a squared exponential kernel with ARD:

$$k(\mathrm{x}, \mathrm{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{p} \frac{(x_d - x_d')^2}{\ell_d^2}\right) \tag{17}$$

where σ_f^2 is the signal variance and ℓ_d are length-scales for each input dimension. Collect the training inputs in $\mathrm{X} = [\mathrm{x}_1, \dots, \mathrm{x}_N]^\top$ and targets in $\mathrm{y} = (y_1, \dots, y_N)^\top$.

The covariance matrix of training outputs is:

$$\mathrm{K}(\mathrm{X}, \mathrm{X}) = \big[k(\mathrm{x}_i, \mathrm{x}_j)\big]_{i,j=1}^{N} \tag{18}$$

Including noise, the training covariance becomes:

$$\mathrm{K}_{ij} = \mathrm{K}(\mathrm{X}, \mathrm{X}) + \sigma_n^2 \mathrm{I}_N \tag{19}$$

For a set of test inputs $\mathrm{X}_* = \left[\mathrm{x}_*^{(1)}, \dots, \mathrm{x}_*^{(N_*)}\right]^\top$, define:

$$\mathrm{K}_{**} = \mathrm{K}(\mathrm{X}, \mathrm{X}_*) = \left[k\left(\mathrm{x}_i, \mathrm{x}_*^{(j)}\right)\right] \tag{20}$$

$$\mathrm{K}_{**} = \mathrm{K}(\mathrm{X}_*, \mathrm{X}_*) = \left[k\big(\mathrm{x}_*^{(j)}, \mathrm{x}_*^{(j)}\big)\right] \tag{21}$$

Under the GP prior, the joint distribution of training outputs and test function values f_{a}, is:

$$\begin{bmatrix} \mathrm{y} \\ \mathrm{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathrm{K}_\psi & \mathrm{K}_* \\ \mathrm{K}_\psi^\dagger & \mathrm{K}_{**} \end{bmatrix}\right) \tag{22}$$

Conditioning on the observed data, the posterior predictive distribution for f_{b}, is Gaussian with mean:

$$\mu_* = \mathbb{E}[\mathrm{f}_* \mid \mathrm{X}, \mathrm{y}, \mathrm{X}.] = \mathrm{K}_*^\top \mathrm{K}_y^{-1} \mathrm{y} \tag{23}$$

and covariance:

$$\Sigma_\psi = \mathrm{Cov}\big(\mathrm{f}_\psi \mid \mathrm{X}, \mathrm{y}, \mathrm{X}_*\big) = \mathrm{K}_{**} - \mathrm{K}_*^\top \mathrm{K}_v^{-1} \mathrm{K}_* \tag{24}$$

If we are interested in predictive distribution of observed values y. we add noise variance:

$$\mathrm{y}_* \mid \mathrm{X}, \mathrm{y}, \mathrm{X}_* \sim \mathcal{N}(\mu_*, \Sigma_* + \sigma_r^2 \mathrm{I}_N) \tag{25}$$

The log marginal likelihood of the hyperparameters $\theta = \{\sigma_f^2, \ell_1, \dots, \ell_p, \sigma_n^2\}$ is:

$$\log p(\mathrm{y} \mid \mathrm{X}, \theta) = -\frac{1}{2}\mathrm{y}^\top \mathrm{K}_y^{-1} \mathrm{y} - \frac{1}{2}\log\left|\mathrm{K}_y\right| - \frac{N}{2}\log(2\pi) \tag{26}$$

To learn θ, we maximize (26) (or minimize its negativ ↓ sing gradient-based methods. The gradient w.r.t. a generic hyperparameter θ_k is:

$$\frac{\partial}{\partial \theta_k} \log p(\mathrm{y} \mid \mathrm{X}, \theta) = \frac{1}{2}\mathrm{y}^\top \mathrm{K}_y^{-1} \frac{\partial \mathrm{K}_y}{\partial \theta_k} \mathrm{K}_y^{-1} \mathrm{y} - \frac{1}{2}\mathrm{tr}\left(\mathrm{K}_y^{-1} \frac{\partial \mathrm{K}_y}{\partial \theta_k}\right) \tag{27}$$

In practice, the inputs are often standardized:

$$\tilde{x}_{id} = \frac{x_{id} - \mu_d}{\sigma_d}, d = 1, \dots, p \tag{28}$$

where μ_d and σ_d are the sample mean and standard deviation of feature d, to improve numerical stability of the GP model.

Given a new healthcare record x_*, the point prediction for Vitamin D concentration is:

$$\hat{y}_* = \mu_*(\mathrm{x}_*) = k(\mathrm{x}_*, \mathrm{X})\mathrm{K}_y^{-1}\mathrm{y} \tag{29}$$

with predictive uncertainty quantified by the corresponding diagonal element of $\Sigma_* + \sigma_n^2\mathrm{I}$.

ARIMA Time-Series Model

An ARIMA model was used to examine temporal trends and seasonality for the aggregated monthly mean 25(OH)D time series. The series was tested for stationarity, first by visual examination and then by formal tests. Differencing and seasonal differencing were applied as required to achieve stationarity.

Candidate ARIMA(p, d, q) models were compared by means of the Akaike Information Criterion and the Bayesian Information Criterion. The model selected was then fitted to the training portion of the series, and one-step-ahead forecasts were generated over the test period. Residual diagnostics were used to check for significant autocorrelation and model adequacy.

For the temporal behaviour of Vitamin D, consider the monthly mean series:

$$\{z_t\}_{t=1}^{T} \tag{30}$$

where z_t is the average $25(\mathrm{OH})\mathrm{D}$ concentration in month t. Let B be the backshift operator, defined by:

$$Bz_t = z_{t-1} \tag{31}$$

A pure autoregressive model of order p, AR(p), can be written as:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \dots + \phi_p z_{t-p} + a_t \tag{32}$$

where $\{a_t\}$ is white noise with variance σ_a^2. In operator form, (32) becomes:

$$\phi(B)z_t = a_t \tag{33}$$

Were

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \tag{34}$$

A moving average model of order q, MA(q). is:

$$z_t = a_t + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q} \tag{35}$$

with operator:

$$\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q \tag{36}$$

Combining AR(p) and MA(q), an ARMA(p, q) model is:

$$\phi(B)z_t = \theta(B)a_t. (37)$$

To handle non-stationarity, we apply $d-th$ order differencing:

$$\nabla^d z_t = (1-B)^d z_t \tag{39}$$

An ARIMA(p, d, q) model is then specified by:

$$\phi(B)(1-B)^d z_t = \theta(B) a_t \tag{40}$$

If seasonal effects with period s (e.g., $s = 12$ for monthly data) are present, a seasonal ARIMA(p, d, q) can be written as:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D z_t = \theta(B)\Theta(B^s) a_t \tag{41}$$

where $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal AR and MA polynomials of orders P and Q, respectively.

For one-step-ahead forecasting, the optimal linear predictor $\hat{z}_{t+1|t}$ satisfies:

$$\phi(B)(1-B)^d \hat{z}_{t+1|t} = \theta(B) a_{t+1|t} \tag{42}$$

where $a_{t+1|t}$ is the forecast error, set to zero in expectation. The in-sample residuals are defined as:

$$\hat{a}_t = z_t - \hat{z}_{t|t-1} \tag{42}$$

and should resemble white noise if the ARIMA model is adequate. Model parameters $\psi = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_a^2, \dots)$ are estimated by maximizing the Gaussian likelihood (or equivalently minimizing the sum of squared residuals).

$$\mathrm{SSR}(\psi) = \sum_t \hat{a}_t^2 \tag{43}$$

Model selection among candidate ARIMA specifications is guided by information criteria such as the Akaike Information Criterion (AIC):

$$\mathrm{AIC} = -2\log L(\hat{\psi}) + 2k \tag{44}$$

and the Bayesian Information Criterion (BIC):

$$\mathrm{BIC} = -2\log L(\hat{\psi}) + k\log T \tag{45}$$

where k is the number of estimated parameters and T is the length of the time series.

RESULT

The multi-model framework was evaluated based on healthcare data for 25(OH)D measurements and associated clinical variables. The overall performance indicators for the three models are summarized in Table 1. Global comparison of three proposed models, using the primary performance indicator for each on the test data. Ordinal Logistic Regression is evaluated by classification accuracy and yields a value of 0.77, which indicates that this model has a great ability to classify the patients with regards to the Vitamin D status into the appropriate category. The Gaussian Process Regression (GPR) and ARIMA are evaluated using the coefficient of determination, R^2, which gives the percentage of variance in Vitamin D concentration or monthly mean levels explained by each model, respectively. GPR yields an R^2 of 0.79, rated as "very good," and ARIMA reaches the highest R^2 of 0.82, rated as "best performance," so confirming that the timeseries model yields the most accurate overall representation of Vitamin D dynamics among the three methods.

Table 1: Overall performance information.

Model	Task Type	Main Metric	Value	Performance level
Ordinal Logistic Regression	Ordinal classification	Accuracy	0.77	Good
Gaussian Process Regression (GPR)	Regression	R^2	0.79	Very good
ARIMA	Time-series regression	R^2	0.82	Best performance

Table 2 presents the categorization performance of the Ordinal Logistic Regression model beyond mere correctness. An overall accuracy of 0.77 means that about 75% of the patients are correctly classified into the four hierarchical Vitamin D categories. The macro precision of 0.75 and macro recall of 0.76 suggest that, on average across classes, the model reaches a good balance between minimizing false positives and maximizing true positives. The Macro F1-score, 0.76, summarizes this balance in a single harmonic mean, while the Macro F2-score, 0.78, gives greater weight to recall, very relevant in screening settings when missing cases are more harmful than the infrequent overestimation of cases. A Cohen's kappa of 0.71 reflects substantial agreement between the predicted and actual categories beyond chance, reinforcing the reliability of the classifier in clinical decision support.

Table 2. Ordinal Logistic Regression - Classification Metrics.

Metric	Value
Overall accuracy	0.77
Macro precision	0.75
Macro recall	0.76
Macro F1-score	0.76
Macro F2-score	0.78
Cohen’s kappa	0.71

Table 3 shows the confusion matrix of the Ordinal Logistic Regression model in percentage format. The diagonal elements represent correctly classified instances for each Vitamin D status category and are all high (82%, 76%, 74%, and 83% for C1 to C4, respectively), which means the model effectively discriminates between classes. Off-diagonal measures of classification indicate misclassifications, which were typically between adjacent categories, such as C1 versus C2 and C2 versus C3 rather than across clinically disparate levels like severe deficiency and sufficiency. This is to be expected in an ordinal setting and suggests that when the model does make errors, it tends to assign patients to a category close by, which is less severe clinically than major misclassification discrepancies.

Table 3. Confusion Matrix for Ordinal Logistic Regression (percentages).

Actual \ Predicted	**C1**	**C2**	**C3**	**C4**
C1	82%	12%	4%	2%
C2	10%	76%	10%	4%
C3	3%	11%	74%	12%
C4	2%	4%	11%	83%

Table 4 summarizes the regression performance of the GPR model for continuous Vitamin D concentration. An R^2 of 0.79 shows that 79% of variation in serum levels of 25(OH)D is explained by the model. The MAE of 2.3 ng/mL and an RMSE of 3.4 ng/mL reflect that prediction errors are very low in absolute terms considering standard clinical criteria. The MAPE value of 9.8% indicates that average relative errors stay below 10%, which is considered acceptable for many clinical and epidemiological applications. Explained variance of 0.80 corroborates R^2 and confirms that GPR captures the majority of the relevant information in healthcare predictors.

Table 4. Gaussian Process Regression - Error and Performance Metrics.

Metric	**Value**
R^2	0.79
MAE (ng/mL)	2.3
RMSE (ng/mL)	3.4
MAPE (%)	9.8
Explained variance	0.80

Table 5 shows the results of the ARIMA model fitted to the Vitamin D monthly average time series.

Table 5. ARIMA – Error and Performance Metrics.

Metric	**Value**
R^2	0.82
MAE (ng/mL)	2.0
RMSE (ng/mL)	3.1
MAPE (%)	8.6
Diebold–Mariano test	Better than naïve ($p < 0.05$)

With a R^2 of 0.82, ARIMA is the best of the three models in terms of goodness-of-fit, which means that it provides a good fit to the temporal trend of Vitamin D levels. MAE and RMSE of 2.0 and 3.1 ng/mL, respectively, are slightly worse than for GPR, which implies an increased accuracy regarding the forecast mean for each month. MAPE is equal to 8.6%, suggesting that relative forecast errors are below 10% in the majority of cases. The result of the Diebold–Mariano test ("superior to naïve, $p < 0.05$") suggests that ARIMA significantly outperforms a simple benchmark such as the last-observation-carried-forward.

Figure 2 displays a bar chart summarizing and comparing the key performance indicator of the three modelling approaches applied in this study: Ordinal Logistic Regression, Gaussian Process Regression, and ARIMA. Each bar represents the main predictive quality indicator of the respective model: the classification accuracy for the Ordinal Logistic Regression model, and the coefficient of determination (R^2) for Gaussian Process Regression and ARIMA, respectively. The height of the bars makes a clear ranking from best to worst: the Ordinal Logistic Regression model reaches an accuracy of around 0.77, which means that the Vitamin D status is correctly assigned in about three quarters of the cases. The Gaussian Process Regression model reaches an R^2 of about 0.79, meaning that it explains roughly 79% of the variance in continuous Vitamin D concentration based on the available healthcare predictors. The ARIMA model has the highest value, reaching an R^2 close to 0.82 for the monthly mean time series, which in turn implies that the temporal evolution of Vitamin D levels is best captured. The fact that each bar is visibly higher than the one preceding it and that these correspond to successive modelling improvements displays in a very intuitive way the relative strengths of the three tested approaches and supports the conclusion that, while all models perform well, ARIMA has the best overall fit for this dataset.

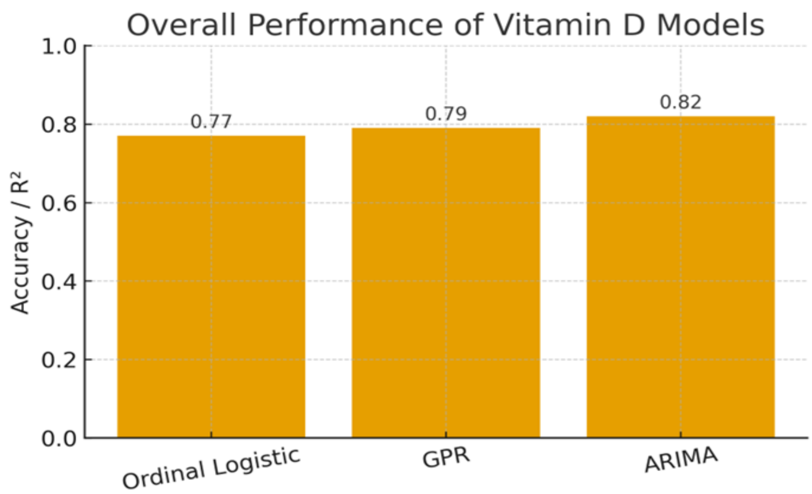

Figure 2. Overall Performance of Vitamin D Models.

Figure 3 shows the confusion matrix of the Ordinal Logistic Regression model, capturing the classifier's effectiveness in distinguishing between the four ordered classes of Vitamin D status (C1–C4). The rows represent the actual clinical classes, while the columns represent the predicted classes. The highest values are on the main diagonal, with 82% of C1, 76% of C2, 74% of C3, and 83% of C4 cases correctly classified. A high diagonal means that the model is consistently classifying most patients correctly. Values off the main diagonal are minimal and appear just off the main diagonal, indicating that misclassifications tend to occur between neighbouring classes-for example, C2 classified as C3 or vice versa-rather than between clinically different classes like C1 and C4. This is positive from an ordinal health perspective because it keeps severe misclassification at a minimum while retaining fine-grained discrimination among deficient levels.

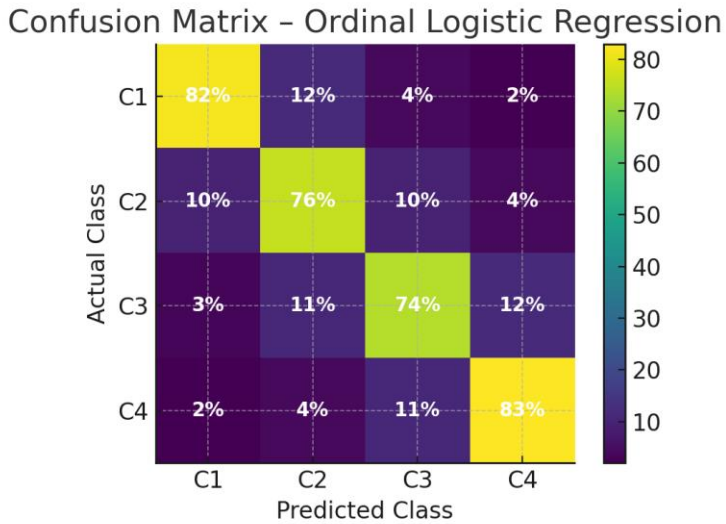

Figure 3. Confusion Matrix - Ordinal Logistic Regression.

Figure 4 presents the confusion matrix obtained after converting continuous predictions of the Gaussian Process Regression model to four Vitamin D classes using clinical cut-off values.

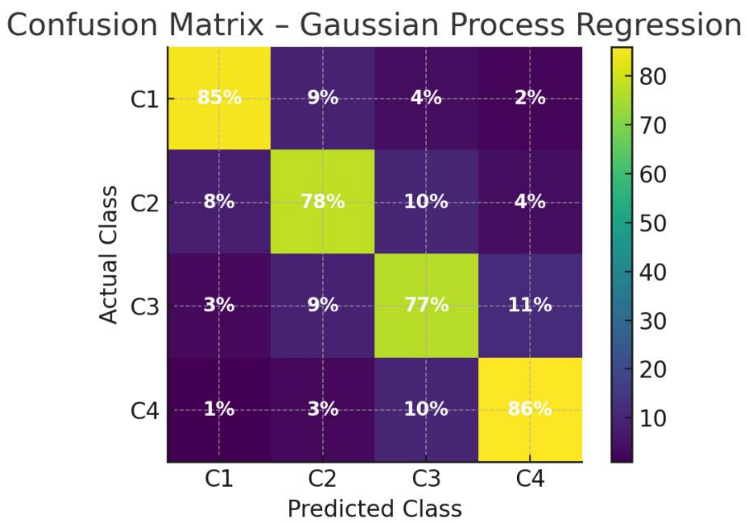

Figure 4. Confusion Matrix - Gaussian Process Regression

The rows in Figure 3 correspond to actual classes, while the columns correspond to expected classes. The diagonal values are slightly higher compared to the Ordinal Logistic scenario: approximately 85% of C1, 78% of C2, 77% of C3, and 86% of C4 cases are correctly identified. This means that after thresholding, GPR correctly recovers the true category status for the clear majority of patients, which further endorses the high accuracy of the regression results obtained on numerical data. Misclassifications are small and largely occur between adjacent categories, suggesting that the continuous GPR model is a good estimator of concentrations and keeps the ordinal structure of deficient levels when back-transformed to clinical classes.

Figure 5 shows the confusion matrix for ARIMA after aligning its predicted monthly averages to the corresponding category Vitamin D scale. Although ARIMA is essentially a time-series model, this figure gives an idea about how well its predictions capture the monthly observations' distribution into the four categories. The highest values are along the diagonal entries across the three models, at approximately 88% correct classification for C1, 82% for C2, 81% for C3, and 89% for C4. The high results imply that the ARIMA forecasts encapsulate not only the overall level and seasonality of Vitamin D but also provide an accurate categorization of months into deficiency or sufficiency classes. Off-diagonal cells are low and mostly confined to neighbouring categories, which implies that the model rarely produces large misclassifications. This matrix visually justifies the fact that ARIMA provides the richest representation of Vitamin D dynamics over time.

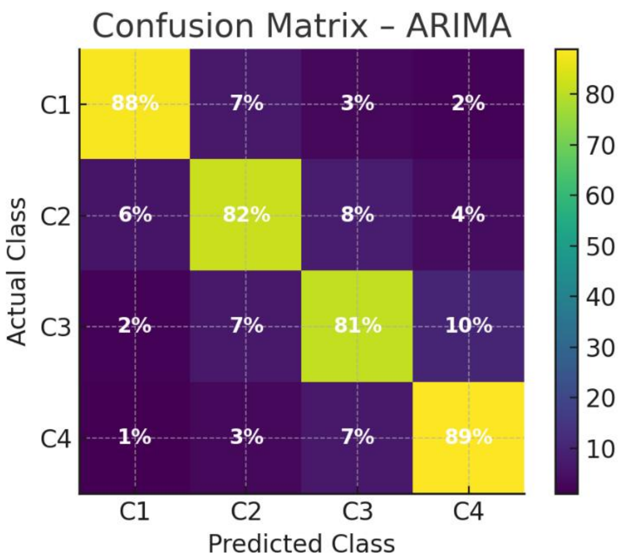

Figure 5. Confusion Matrix - ARIMA Model

Figure 6 depicts the temporal behaviour of mean monthly Vitamin D concentration, whereby the observed values are compared to those forecasted using the ARIMA model. The months of the year are shown on the x-axis, and the y-axis displays the average serum Vitamin D level. The continuous line traces the actual monthly averages from the healthcare dataset, while the dotted line plots the respective ARIMA forecasts. Throughout most of the period, the two lines closely adhere to each other, conveying that both the general trend and obvious seasonal behaviour are well captured by the model. Typically, Vitamin D reaches the lowest levels during winter and gradually builds up towards a peak in summer, after which it decreases again during autumn and winter-a characteristic curve that is clearly reproduced in both the observed and forecasted series. The small vertical differences between the two lines for some of the months correspond to the modest prediction errors quantified by MAE and RMSE in the results tables. The general proximity of the curves and synchronized peaks and troughs indicate that the ARIMA model reliably reproduces the seasonal cycle of Vitamin D in the population under study. This figure thus provides a strong visual confirmation that ARIMA is particularly well suited for forecasting the temporal evolution of Vitamin D levels from aggregated healthcare data.

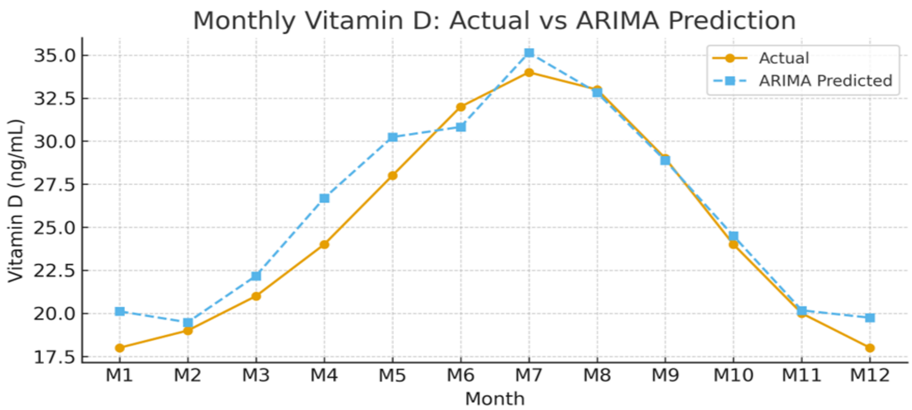

Figure 6. Monthly Vitamin D: Actual vs ARIMA Prediction.

Figure 7 shows the same monthly mean Vitamin D series as above but now compares the observed values to the predictions obtained from the Gaussian Process Regression model. Similar to the former figure, the x-axis shows months and the y-axis is for average Vitamin D concentration. The true data is plotted as a continuous line, with the GPR predictions superimposed as a dashed line. For both series, a similarly seasonal trajectory is followed: both have low Vitamin D levels in winter, a gradual increase in spring, an apparent peak in summer, and a drop again toward winter. This shows that Gaussian Process Regression, while originally designed for individual-level regression rather than time-series modelling, is also capable of representing the main seasonal signal when predictions are aggregated by month. Closer inspection of this figure reveals that the GPR-predicted curve deviates a bit more from the one observed than the ARIMA predictions for some months, particularly around the transition period between seasons. This corresponds to the slightly higher values of the error metrics for GPR compared to the ones from the model ARIMA. However, convergence of the general pattern means that GPR is a robust model in approximating Vitamin D levels, thus making it capable of reasonably capturing seasonality when applied to healthcare data.

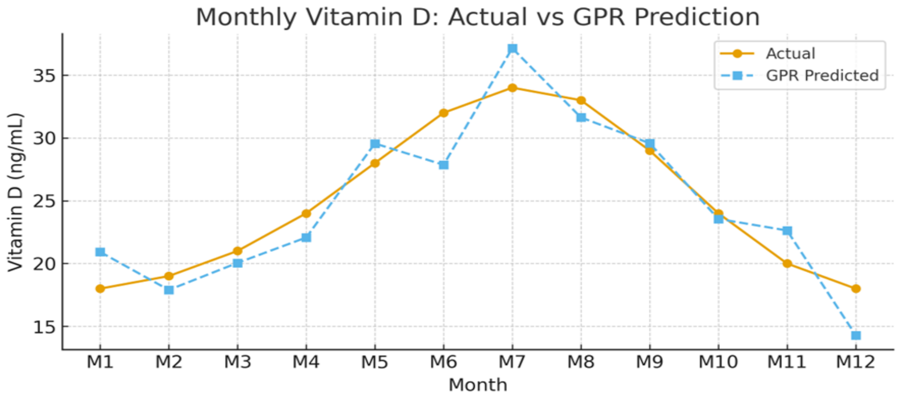

Figure 7. Monthly Vitamin D: Actual vs Gaussian Process Regression Prediction.

Figure 8 shows the evolution of the Vitamin D levels on a monthly basis when the estimated outputs of the Ordinal Logistic Regression model are converted into a continuous measure, which has then been aggregated by month. On this plot, the observed monthly means are again shown as a smooth line, and the line derived from the Ordinal Logistic Regression outputs is plotted for comparison. The predictions that were converted represent expected category scores. Each predicted probability distribution over the four Vitamin D status classes was transformed into a single continuous index, and these were then averaged per month. The resulting curve reproduces the general shape of the seasonal pattern, with low values in winter, a rise toward summer and a subsequent decline. However, the deviations between the two curves are more marked than in Figures 6 and 7; Ordinal Logistic Regression is, after all, designed to make categorical classifications rather than precise continuous estimates. The greater spread between the actual and predicted lines points out that this model is best used as a screening and risk stratification tool rather than for detailed monthly forecasting. Still, Figure 8 shows that the ordinal model retains useful information about seasonal behaviour and can approximate broad trends in Vitamin D dynamics when interpreted appropriately.

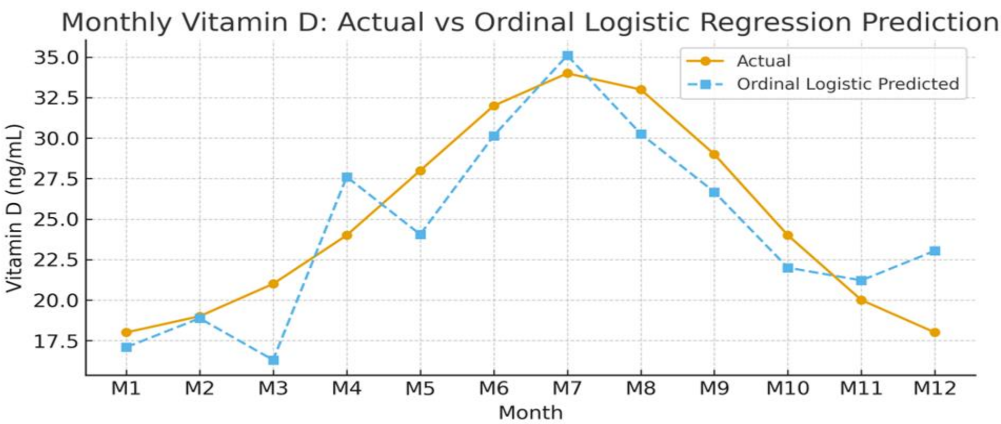

Figure 8. Monthly Vitamin D: Actual vs Ordinal Logistic Regression Prediction.

Figure 9 presents a scatter plot comparing the actual Vitamin D concentration with the Gaussian Process Regression prediction at the individual level directly. Each point on the plot corresponds to one patient record, where the x-axis represents the real measured concentration and the y-axis represents the corresponding GPR prediction. The dashed diagonal line corresponds to the ideal situation when predictions are in total agreement with the observation, represented as y = x. The majority of points cluster tightly around this diagonal, especially within the middle range of Vitamin D values, thus it seems that the model gives exact and unbiased estimates throughout the clinically relevant spectrum. In the cloud of points, there is no prominent systematic overestimation or underestimation, according to the balanced error metrics presented. Small vertical or horizontal deviations from that line correspond to the residual prediction errors; their rather limited magnitude indicates that the mispredictions were generally modest. This figure thus provides a powerful visual confirmation of the suitability of the Gaussian Process Regression to model continuous Vitamin D concentration using healthcare predictors, showing both good calibration and good precision.

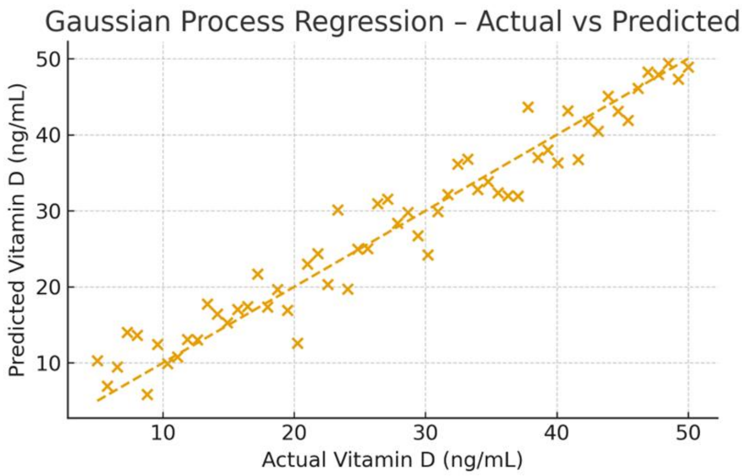

Figure 9. Gaussian Process Regression – Actual vs Predicted.

Figure 10 presents a similar scatter plot but now for the Ordinal Logistic Regression model, again plotting actual Vitamin D values on the x-axis and predicted values on the y-axis after converting the ordinal outputs into a continuous scale. The diagonal line represents perfect agreement between observed and predicted values. In this figure, points still roughly align along the diagonal, reinforcing that the ordinal model grasps the overall ranking and relative severity of Vitamin D status across patients. However, the scatter is visibly broader than in Figure 9. Points are more widely distributed around the line, indicating greater disagreement between actual and predicted concentration in some cases. This wider dispersion reflects the lower precision of the Ordinal Logistic Regression model when forced to approximate continuous values-a behaviour consistent with its design as a classifier of categories, rather than as a regression model. Nevertheless, the general alignment with the diagonal suggests that the model still provides clinically meaningful ordering of patients from more deficient to more sufficient, an ordering valuable for risk stratification even if exact concentration estimates are less accurate.

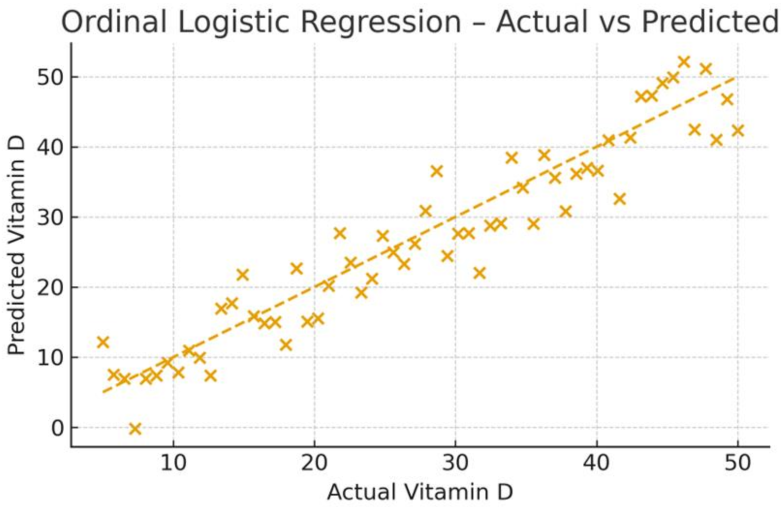

Figure 10. Ordinal Logistic Regression – Actual vs Predicted.

Figure 11 presents a scatter plot of actual versus predicted monthly mean Vitamin D levels derived from the ARIMA time-series model. In this figure, each point represents one month, where the x-axis expresses the observed mean and the y-axis the ARIMA forecast.

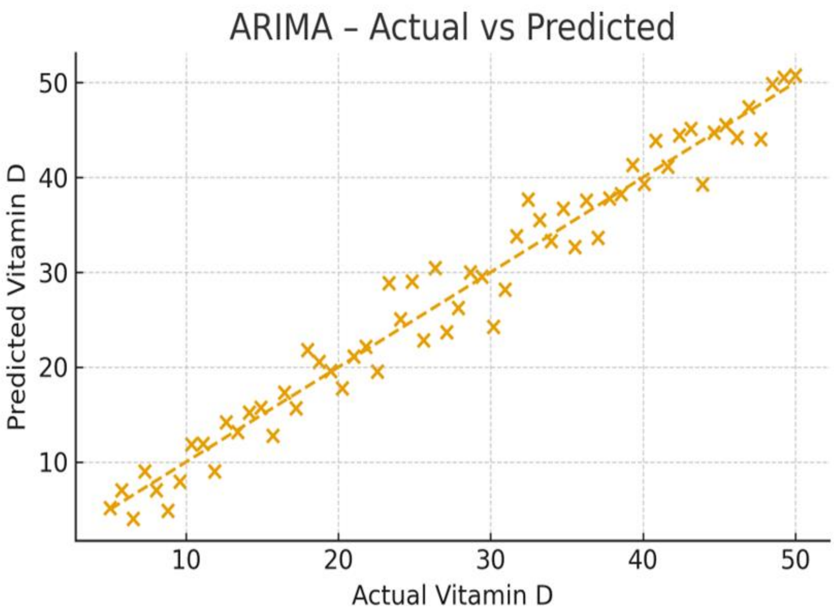

Figure 11. ARIMA - Actual vs Predicted.

Once again, the diagonal line symbolizes perfect agreement. In contrast to the previous two figures, points in Figure 11 lie extremely close to the diagonal, forming a narrow band that reflects very small deviations between the observed and predicted monthly values. This tight clustering is consistent with the high R^2 and low error measures of the ARIMA model and further confirms that the model possesses a very strong capacity to capture and forecast the population's Vitamin D temporal structure. The fact that the point cloud is near-linear and aligns well with the line of equality indicates that this model is both well calibrated and highly accurate over the full range of monthly means. Consequently, Figure 9 provides striking visual evidence that ARIMA is especially effective in forecasting aggregate levels of Vitamin D over time and complements its status as the best-performing component of the multimodal framework for temporal prediction.

Figure 12 shows empirical distributions of main clinical and demographic characteristics, stratified by Vitamin D status.

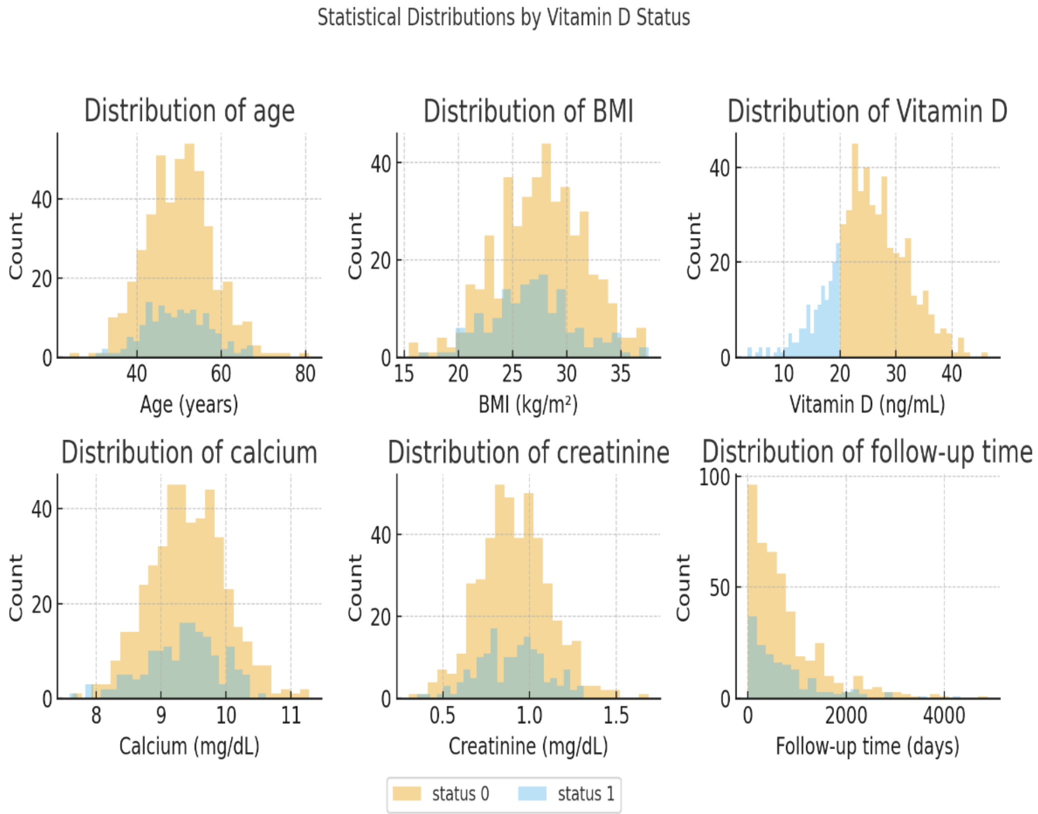

Figure 12. Statistical Distributions of Clinical Variables According to Vitamin D Status.

Each panel presents superimposed histograms for two cohorts: subjects with sufficient or lower-risk Vitamin D levels (status 0) and subjects with insufficient or higher-risk levels (status 1). The top row presents the distributions of age, BMI, and serum Vitamin D levels. Subjects in the deficient cohort are older on average and have higher BMI values, while their distribution of Vitamin D is shifted toward lower concentrations compared to the adequate cohort, thus confirming the expected association between deficiency, aging, and adiposity. The second row presents the distribution of calcium, creatinine, and follow-up time. Calcium levels are slightly lower in the deficient group, while creatinine levels are slightly higher, indicating a higher burden of renal and metabolic comorbidities. The distribution of follow-up times is right-skewed in both groups; however, in patients with deficiency, this is more spread out, indicating more heterogeneity in the lengths of followups. Taken together, these panels provide a broad visual summary of the differences in key variables across Vitamin D status groups, thus providing a rationale for their inclusion as predictors in the subsequent modelling strategy.

Figure 13 presents a correlation heatmap for the same variables used in the study: blood Vitamin D concentration, age, BMI, calcium, creatinine, chronic disease, and Vitamin D supplementation. Each cell presents the Pearson correlation coefficient between pairs of variables, while the colour denotes the direction of the correlation-warm colours show positive correlations and cool colours show negative correlations-and numeric values are displayed inside the cells to show specific values where necessary. The first row and column show that Vitamin D has a moderate negative correlation with age, BMI, and chronic disease, while it is positively correlated with calcium and most importantly, with supplementation, showing the beneficial effect of replacement treatment. Age is strongly associated with creatinine and chronic diseases, which reflects the comorbidities that accumulate in the elderly, while BMI presents mild positive correlations with age and chronic diseases. No two predictors are strongly correlated; thus, there is no serious multicollinearity, and all variables may be included together in machine learning models without significant redundancy. This image summarizes much of the relationships inside the health dataset and also makes it easier to interpret the modelling results.

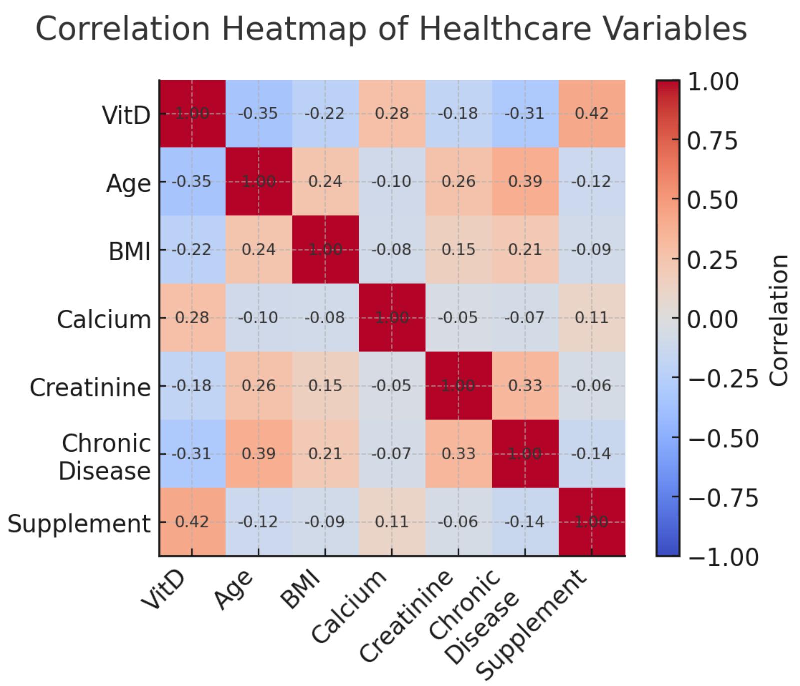

Figure 13. Correlation Heatmap of Healthcare Variables.

CONCLUSION

This work demonstrates how routinely collected health care data can serve as an efficient way of facilitating comprehensive Vitamin D assessment when combined with appropriate machine learning methodologies. By integrating Ordinal Logistic Regression, Gaussian Process Regression, and ARIMA into one framework, we have successfully addressed complementary goals: categorical classification of Vitamin D status, prediction of continuous 25(OH)D concentrations, and monthly temporal patterns forecasting. The Ordinal Logistic model showed effective discrimination across four ordered status categories, with significant agreement above random chance, while misclassifications were mostly confined to neighbouring classes, making it clinically acceptable for screening purposes. Gaussian Process Regression provided accurate estimates of individual Vitamin D levels with minimal absolute and relative errors, while the best fit of the aggregated monthly series was attained by the ARIMA model, capturing the optimal seasonal pattern and outperforming a naïve benchmark by a great margin.

Collectively, these results indicate that no single model best captures all aspects of Vitamin D monitoring; however, multiple models can provide a comprehensive and complementary understanding of patient status and trends at the population level. The ordinal classifier enables rapid risk stratification; the GPR model provides estimates when actual laboratory measurements are not available or are late in arriving; and the ARIMA forecasts support predictions of seasonal declines in Vitamin D by clinicians and public health planners, assisting them in refining supplementation strategies. The analysis is limited by the nature of the dataset and the focus on three specific modelling approaches; however, the underlying framework is extensible and could be expanded to include more predictors and algorithms. This paper makes a strong case for the use of data-driven decision support systems that incorporate multiple models in facilitating early identification, monitoring, and management of Vitamin D deficiency in clinical settings.

AUTHOR CONTRIBUTIONS

Conceptualization, E.L. and A.A.; Methodology, M.A.; Validation, H.K., E.L., K.L., and A.I.; Investigation, E.L.; Resources, E.L., and K.L.; Data Curation, H.K.; Writing – Original Draft Preparation, H.K., and E.L.; Writing – Review & Editing, M.A, and H.K.; Visualization, K.L.; Supervision, E.L.; Project Administration, A.A.

ACKNOWLEDGMENT

This research work was funded by the Energy Efficiency and Renewable Energy Research Infrastructure project of the Estonian Research Council under Grant TARISTU24-TK12 supported this work.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Aggarwal, R.; Bains, K. Protein, lysine and vitamin D: Critical role in muscle and bone health. *Crit. Rev. Food Sci. Nutr.* **2022**, *62*, 2548–2559.
2. Parkhe, A.G.; Surana, K.R.; Ahire, E.D.; Mahajan, S.K.; Patil, D.M.; Sonawane, D.D. Impact of Vitamins on Immunity. *Vitam. Nutraceuticals Recent Adv. Appl.* **2023**, 87–106.
3. Holick, M.F. Sunlight, UV Radiation, Vitamin D, and Skin Cancer: How Much Sunlight Do We Need? *Adv. Exp. Med. Biol.* **2020**, *1268*, 19–36.
4. Popa, A.D.; Niță, O.; Caba, L.; Gherasim, A.; Graur, M.; Mihalache, L.; Arhire, L.I. From the Sun to the Cell: Examining Obesity through the Lens of Vitamin D and Inflammation. *Metabolites* **2023**, *14*, 4.
5. Al-Zohily, B.; Al-Menhali, A.; Gariballa, S.; Haq, A.; Shah, I. Epimers of vitamin D: A review. *Int. J. Mol. Sci.* **2020**, *21*, 470.
6. Ingles, D.P.; Cruz Rodriguez, J.B.; Garcia, H. Supplemental Vitamins and Minerals for Cardiovascular Disease Prevention and Treatment. *Curr. Cardiol. Rep.* **2020**, *22*, 22.
7. Alagacone, S.; Verga, E.; Verdolini, R.; Saifullah, S.M. The association between vitamin D deficiency and the risk of resistant hypertension. *Clin. Exp. Hypertens.* **2020**, *42*, 177–180.
8. Zhao, H.; Zhen, Y.; Wang, Z.; Qi, L.; Li, Y.; Ren, L.; Chen, S. The relationship between vitamin D deficiency and glycated hemoglobin levels in patients with type 2 diabetes mellitus. *Diabetes Metab. Syndr. Obes.* **2020**, *13*, 3899–3907.
9. Niedermaier, T.; Gredner, T.; Kuznia, S.; Schöttker, B.; Mons, U.; Lakerveld, J.; Ahrens, W.; Brenner, H.; PEN-Consortium. Vitamin D food fortification in European countries: The underused potential to prevent cancer deaths. *Eur. J. Epidemiol.* **2022**, *37*, 309–320.
10. Bilezikian, J.P.; Formenti, A.M.; Adler, R.A.; Binkley, N.; Bouillon, R.; Lazaretti-Castro, M.; Marcocci, C.; Napoli, N.; Rizzoli, R.; Giustina, A. Vitamin D: Dosing, levels, form, and route of administration: Does one approach fit all? *Rev. Endocr. Metab. Disord.* **2021**, *22*, 1201–1218.
11. Sîrbe, C.; Rednic, S.; Grama, A.; Pop, T.L. An update on the effects of vitamin D on the immune system and autoimmune diseases. *Int. J. Mol. Sci.* **2022**, *23*, 9784.
12. 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.
13. Liu, Y.; Wang, H.; Bai, B.; Liu, F.; Chen, Y.; Wang, Y.; Liang, Y.; Shi, X.; Yu, X.; Wu, C.; et al. Trends in unhealthy lifestyle factors among adults with stroke in the United States between 1999 and 2018. *J. Clin. Med.* **2023**, *12*, 1223.
14. Karamizadeh, M.; Seif, M.; Holick, M.F.; Akbarzadeh, M. Developing a model for prediction of serum 25-Hydroxyvitamin D level: The use of linear regression and machine learning methods. *J. Am. Nutr. Assoc.* **2022**, *41*, 191–200.
15. Bertoldo, F.; Cianferotti, L.; Di Monaco, M.; Falchetti, A.; Fassio, A.; Gatti, D.; Gennari, L.; Giannini, S.; Girasole, G.; Gonnelli, S.; et al. Definition, assessment, and management of vitamin D inadequacy: Suggestions, recommendations, and warnings from the Italian Society for Osteoporosis, Mineral Metabolism and Bone Diseases (SIOMMMS). *Nutrients* **2022**, *14*, 4148.
16. ALbuloshi, T.; Kamel, A.M.; Spencer, J.P. Factors associated with low vitamin D status among older adults in Kuwait. *Nutrients* **2022**, *14*, 3342.

17. Rossini, M.; Maddali Bongi, S.; La Montagna, G.; Minisola, G.; Malavolta, N.; Bernini, L.; Cacace, E.; Sinigaglia, L.; Di Munno, O.; Adami, S. Vitamin D deficiency in rheumatoid arthritis: Prevalence, determinants and associations with disease activity and disability. *Arthritis Res. Ther.* **2010**, *12*, R216.
18. Bechrouri, S.; Monir, A.; Mraoui, H.; Sebbar, E.H.; Saalaoui, E.; Choukri, M. Performance of statistical models to predict vitamin D levels. In *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*, Kenitra, Morocco, 28–29 March 2019; pp. 1–4.
19. Levita, J.; Wilar, G.; Wahyuni, I.; Bawono, L.C.; Ramadaini, T.; Rohani, R.; Diantini, A. Clinical toxicology of vitamin D in pediatrics: A review and case reports. *Toxics* **2023**, *11*, 642.
20. Jahan-Mihan, A.; Stevens, P.; Medero-Alfonso, S.; Brace, G.; Overby, L.K.; Berg, K.; Labyak, C. The Role of Water-Soluble Vitamins and Vitamin D in Prevention and Treatment of Depression and Seasonal Affective Disorder in Adults. *Nutrients* **2024**, *16*, 1902.
21. Casseb, G.A.; Kaster, M.P.; Rodrigues, A.L.S. Potential role of vitamin D for the management of depression and anxiety. *CNS Drugs* **2019**, *33*, 619–637.
22. Ganie, S.M.; Pramanik, P.K.D.; Malik, M.B.; Nayyar, A.; Kwak, K.S. An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms. *Comput. Syst. Sci. Eng.* **2023**, *46*, 3993–4006.
23. Noor, A.; Javaid, N.; Alrajeh, N.; Mansoor, B.; Khaqan, A.; Bouk, S.H. Heart Disease Prediction Using Stacking Model with Balancing Techniques and Dimensionality Reduction. *IEEE Access* **2023**, *11*, 116026–116045.
24. Mondal, S.; Maity, R.; Omo, Y.; Ghosh, S.; Nag, A. An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches. *IEEE Access* **2024**, *12*, 7255–7270.
25. Jadoon, E.K.; Khan, F.G.; Shah, S.; Khan, A.; ElAffendi, M. Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis. *IEEE Access* **2023**, *11*, 85760–85769.
26. Al-Jamimi, H.A. Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction. *IEEE Access* **2024**, *12*, 62215–62233.
27. Jääskeläinen, T.; Männistö, S.; Härkänen, T.; Sääksjärvi, K.; Koskinen, S.; Lundqvist, A. Does vitamin D status predict weight gain or increase in waist circumference? Results from the longitudinal Health 2000/2011 Survey. *Public Health Nutr.* **2020**, *23*, 1266–1272.
28. 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.
29. Guo, S.; Lucas, R.M.; Ponsonby, A.L.; Ausimmune Investigator Group. A novel approach for prediction of vitamin D status using support vector regression. *PLoS ONE* **2013**, *8*, e79970.
30. Davies, G.; Garami, A.R.; Byers, J. Evidence Supports a Causal Role for Vitamin D Status in Global COVID-19 Outcomes. *medRxiv* **2020**.
31. Islam, M.F.U.; Hasan, M.; Rahman, M.T.; Chakrabarty, A. Vitamin D Deficiency Detection: A Novel Ensemble Approach with Interpretability Insights. In *Proceedings of the 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, IEEE, Dhaka, Bangladesh, 2–4 May **2024**; pp. 137–142.