Vitamin Deficiency Disease Prediction Model

John Pierre | ITAI 1371 | Dr. Sina Nazifi

Topic area: Health/fitness

Dataset: Kaggle – Vitamin Deficiency Disease Prediction Dataset

Abstract:

Vitamin and nutritional deficiencies affect millions of people worldwide, yet they often go undetected until symptoms become severe. This problem strains both patients and healthcare providers who rely on costly lab work and clinical intuition for diagnosis. Early, data-driven identification of at-risk individuals could meaningfully reduce this burden, making accurate predictive modeling in this space both timely and clinically relevant. Researchers have already begun tackling this challenge through machine learning; one study published in “Frontiers in Endocrinology” demonstrated that XGBoost outperformed several other models in predicting vitamin D deficiency using NHANES survey data (2001-2018), while a separate team explored a CNN-based web application capable of identifying deficiencies from user-uploaded images — though challenges around data availability, model interpretability, and generalization remain. Industry leaders are pushing the frontier further: Google has developed multimodal health models like MedGemma and a Personal Health LLM built on Gemini, NVIDIA's Clara platform provides researchers with GPU-accelerated tools for medical imaging and drug discovery. Microsoft's InnerEye project applies deep learning to medical image analysis to improve clinical workflows. For this course project, I will build a multi-class disease classification model using the Vitamin Deficiency Disease Prediction Dataset from Kaggle, training and comparing a Random Forest Classifier and XGBoost model across 34 patient features. These features include but are not limited to age, BMI, dietary habits, and symptom data. The goal is to predict which of six deficiency-related disease categories a patient belongs to, evaluated using accuracy, F1-score, ROC-AUC, and per-class precision-recall metrics.

Potential Use Cases:

= Interest/focus

- **Multi-class Disease Classification:** Predict which deficiency disease a patient has based on symptoms and lab values
- **Early Detection Models:** Identify at-risk individuals before severe deficiency develops
- **Feature Importance Analysis:** Understand which factors most strongly predict nutritional diseases
- **Symptom-Disease Correlation:** Map symptom patterns to specific deficiency types
- **Dietary Risk Assessment:** Evaluate how diet types correlate with deficiency risks
- **Public Health Modeling:** Identify vulnerable populations (vegans, low-income, low sun exposure)

- **Clinical Decision Support:** Assist healthcare providers in differential diagnosis

Classification Tasks:

- Multi-class classification (6 disease categories)
- Binary classification (Healthy vs. Diseased)
- Multi-label classification (multiple simultaneous deficiencies)
- Individual symptom prediction

What the Best Work is Doing:

1. Machine learning-based prediction of vitamin D deficiency: NHANES 2001-2018
    a. **What are they trying to solve?**
        i. In this paper the researchers constructed several ML models to predict the risk of vitamin D deficiency.
    b. **What data did they use?**
        i. The National Health and Nutrition Examination Survey 2001-2018 dataset was used for their analysis with a 70:30 split for training and validation
    c. **What ML method did they use?**
        i. GBM, LR, NNet, RF, SVM, XGBoost methods were used and evaluated.
    d. **What's one important result?**
        i. The researchers concluded that the XGBoost-based prediction tool was highly accurate in predicting the risk of vitamin D deficiency in community populations. That is to say that it performed the best.
2. Vitamin Deficiency Detection Using Machine Learning and Deep Learning Algorithms
    a. **What are they trying to solve?**
        i. The authors propose an AI-driven web application that uses machine learning to predict potential vitamin deficiencies based on user-uploaded images and inputs.
    b. **What data did they use?**
        i. Although no formal data set was used, the authors mention several forms of data that could be used, such as medical records, dietary intake, and clinical symptoms.
    c. **What ML method did they use?**
        i. This team used a pre-trained Convolutional Neural Network (CNN), specifically a ResNet-50 model.
    d. **What's one important result?**
        i. One important conclusion was that despite the potential that these models offer there are still challenges that remain (data availability, interpretability of deep models, and model generalization)

A Dive Into 3 Industry Leaders:

Source: Top 10 AI Healthcare Companies – AI Magazine

Google

Sources: Google Health, Google PH-LLM

1. What product or technology are they building?
    a. Google continues to develop AI driven solutions to analyze patient data and health records to make predictions and treatment recommendations.
2. How does ML help them?
    a. Specifically, Google has pioneered technologies such as “MedGemma” and “TxGemma”. Both being AI models with the former providing multimodal medical text and image comprehension and the latter being a collection of open models utilized in the development of therapeutics.
3. What is one interesting detail you learned?
    a. I was particularly intrigued by Google’s “PH-LLM” or “Personal Health LLM”. This is a fine-tuned version of Gemini that provides valuable health insights and recommendations.

NVIDIA

Source: NVIDIA Clara

4. What product or technology are they building?
    a. NVIDIA has a suite specifically for healthcare developers and researchers called “Clara”. Some tools provided include Parabricks, Viz, MONAI, and BioNeMo.
5. How does ML help them?
    a. Although NVIDIA doesn’t directly provide patient care, the aforementioned tools allows researchers and developers to improve healthcare delivery and drug discovery.
6. What is one interesting detail you learned?
    a. One interesting detail I found was that NVIDIA’s DGX systems can help train models to detect cancer in mammograms.

Microsoft

Source: Project InnerEye – Democratizing Medical Imaging AI

7. What product or technology are they building?
    a. Microsoft’s Cloud for Healthcare is a collection of cloud services catered specifically for the health sector.
8. How does ML help them?
    a. Many tools are utilized such as Azure and a research project called “InnerEye” (from Microsoft Health Futures), an AI tool to analyze medical images. InnerEye employs CNNs.
9. What is one interesting detail you learned?
    a. An interesting detail for me was learning about a project called “OSAIRIS” which uses open-source software from InnerEye and Azure Machine Learning. It helps

doctors prepare scans and reduce wait time for patients, allowing them to go from referral to starting treatment quicker.

Questions to be Answered for this Project:

1. **What dataset will you use (or what dataset type)?**
    a. I will be using the Vitamin Deficiency Disease Prediction Dataset from Kaggle (Link at top of doc)
2. **What is your input (features)?**
    a. There are a total of 34 features in this dataset. With it being an exhaustive list, I will only list a few here (gender, bmi, age, smoking_status, etc.). Please reference the dataset for the full list of features.
3. **What is your output (target)?**
    a. Multi-class Disease Classification. The target is to predict which deficiency disease a patient has based on symptoms and lab values
4. **What is your algorithm (Classification, Regression, etc.)?**
    a. I will try both Classification (RFC) and XGBoost.
5. **Which models will you try?**
    a. I will try RFC, but I am particularly interested in XGBoost for it's accuracy.

Recommended Algorithms:

- Random Forest Classifier
- XGBoost / LightGBM
- Neural Networks
- Support Vector Machines
- Logistic Regression (baseline)

Evaluation Metrics:

- Accuracy
- F1-Score (macro/weighted)
- ROC-AUC
- Confusion Matrix Analysis
- Precision-Recall for each disease class