Detecting Vitamin A Deficiency in Schoolchildren Using an Enhanced Explainable Machine Learning Model

Diaa Addeen Abuhani
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
b00086137@aus.edu

Jowaria Khan
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
g00084343@aus.edu

Hana Sulieman
Mathematics and Statistics
American University of Sharjah
Sharjah, United Arab Emirates
hsulieman@aus.edu

2023 Advances in Science and Engineering Technology International Conferences (ASET) | 978-1-6654-5474-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASET56582.2023.10180556

***Abstract*— The most prevalent avoidable cause of vision impairments in children worldwide is vitamin A deficiency. In most cases, deficiencies can be detected through blood tests. However, blood tests are less accessible and of high cost in underdeveloped countries in Africa and Southeast Asia which hinders the efforts of detecting Vitamin A deficiency soon enough to prevent further complications. With the development of machine learning and deep learning in risk-averse industries like healthcare and the expansion of electronic health records, there is a potential to use these techniques in order to arrange for a more accessible substitute to blood tests. In this study, a variety of machine learning techniques are applied to a sparse dataset of ocular symptoms and diagnoses that was obtained from Maradi, Niger, during routine eye exams carried out in a school environment. The goal is to provide an affordable, accessible, and effective clinical screening system for Vitamin A deficiency in children using solely existing health records. The LGB model achieved the best accuracy: 84.4% with a sensitivity of 81.9%, and a specificity of 84.7 which outperformers results on the same dataset recently published [1] by almost 10% in terms of accuracy, specificity, and F1-score.**

Keywords—Machine Learning, Target Encoding, Vitamin A, Ophthalmology, Electronic Health Records, XAI

I. Introduction

Vitamin A Deficiency (VAD) is one of the most common nutritional concerns in low-income societies. Deficiency for a sufficient period can lead to numerous disorders such as Xerophthalmia which is the main cause of childhood blindness [2]. In fact, the World Health Organization (WHO) estimates that 250,000 – 500,000 children become blind because of vitamin A deficiency on a yearly basis [3]. VAD is usually detected through clinical assessments of eye signs or biochemically determined by measuring the concentrations of retinol in serum taken through blood tests. However, access to quality-assured laboratory diagnosis in developing countries is difficult and results in rapid delays, inaccurate diagnosis and inadequate treatment with consequences affecting patient safety [4].

The medical and health sciences fields have shown growing interest in employing Machine Learning (ML) and Deep Learning (DL) techniques to enhance clinical decision making. Nonetheless, a highly accurate ML or DL model alone does not provide enough information for clinical assessments. The issue arises because healthcare is of risk-sensitive applications and cannot blindly trust AI recommendations and insights. Most AI systems operate as black boxes which do not provide further information on how and why a certain decision was made. As a result, [5] introduced the field of Explanatory Artificial Intelligence (XAI) which aims to provide intelligible explanations to the end user. In this paper, we apply two XAI methods to demonstrate the usability of explainable ML or DL classification of VAD in schoolchildren. With the presence of Electronic Health Records (EHS) and the increased utilization of ML and DL in clinical assessments, we believe that there is a potential of early VAD detection through symptomatologic analysis and interpretable ML and DL models.

The following is a list of the study's contributions.

1) Enhace the state-of-the-art machine learning methods that utilizes EHS and refractive error diagnoses obtained from routine checkups of school children in low-income countries.
2) As an alternative to using two separate XAI techniques, offer an understandable Vitamin A deficiency screening solution.

The rest of the paper is organized as follows: First, the background section gives an overview on the problem at hand and previous works to tackle it. Then we describe the dataset used in this study, pre-processing techniques, machine learning models, and XAI methods utilized to interpret our results. Later, the results section summarizes the outcomes of our study in comparison to state-of-the-art results. The discussion section provides a comprehensive analysis of the results and outlines meaningful relationships obtained from the different XAI methods.

II. Background

Although artificial intelligence (AI) is now a crucial part of clinical and remote healthcare applications, the best AI systems are frequently too complicated to be self-explanatory. When dealing with sensitive and private health data, XAI techniques—which are specified to reveal the logic behind the system's predictions and decisions—become even more important [6]. As it gives comprehensive information on every algorithmic step believed to be trustworthy within the medical domain, practitioners, and specialists, XAI is required to be reviewed with the medical domain progression. The three XAI stages can be given as (i) an explainable building process for aiding acceptance, (ii) explainable decisions for enabling trust with users and administrators, and (iii) explainable decisions for the interoperability with business logic [7]. A recent survey on medical XAI places a singular emphasis on interpretability [8]. The requirement for improved interpretability by the algorithm is due to the high level of accountability and transparency found in the medical industry. Both perceptual and mathematical structures fall under the many interpretability categories mentioned here. The saliency maps such as LIME, CAM, Layer-wise Relevance Propagation (LRP), and others can be used to

examine the perceptive interpretability, which is primarily visual evidence. Furthermore, SHAP which is a state-of-the-art XAI method is also incorporated into our study and is quite suitable for our tabular dataset since it considers prediction task as a game, features as players, and coalitions as all possible feature subsets. Applying SHAP has a multitude of advantages. To start, global explanations can be obtained by aggregating local explanations. Global explanations are also more trustworthy than those obtained by the majority of feature attribution methods due to the axiomatic assumptions built into the SHAP theoretical foundations. Finally, SHAP provides a variety of algorithmic implementations to describe any model [9].

III. Materials and Methods

A. Dataset

The data used for this study was collected through mandated annual eye checkups for school children aged between 6 and 15 years old in the city of Maradi, Niger. The dataset has 86,216 records and 125 variables. Patient-specific data was removed, including name, ID, school name and family occupation. In addition, records of missing data and high redundancy were also neglected. The final dataset used for model training contained 18,423 unique patient records. The final set of predictor variables chosen contained 24 categorical predictor and were identical to those included in [1] for comparison purposes. The variables include the following: presence/absence of itching, blepharitis, conjunctivitis, cataracts, blunt force trauma, ptosis, phthisis bulbi, myopia, hyperopia, astigmatism, exotropia, esotropia, alternative squint, amblyopia, nystagmus, megalocornea, and opaque cornea disorders in addition to age and gender with the categorical target variable being the presence/absence of VAD. Age was divided into three groups as follows: 6-8, 9-11, 12-15. Random undersampling was used to balance the data, which resulted in equal distributions of 1450 samples in each class for a total of 2900 cases.

B. Target Encoding

Target encoding is the process of replacing categorical values with the corresponding mean of the target variable. While other encoding techniques exist for categorical variables such as binary encoding, label encoding, and one-hot encoding, target encoding allows us to avoid generating high number of features while keeping a consistent dimensionality of the original data. Further, recent studies [10] suggest that regularized target encoding performed better than traditional strategies with a constantly superior performance across traditional AI algorithms, perhaps since target encoding the predictors picks up values for the predictors that can better explain the target variable which is VAD in this case. The primary difference between our results and the state-of-the-art results [1] lies in the target encoding of predictor variables.

C. Models

In this work, both conventional and ensemble machine learning models were employed. Specifically, K-Nearest Neighbors (KNN), Light Gradient Boosting (LGB), Logistic Regression (LR), Extreme Gradient Boosting (XGB), Categorical Boosting (CB), Random Forest (RF), and Support Vector Classifier (SVC). Tabular Prior-Data Fitted Network (TabPFN) transformer was also investigated in this study. Most of these models were chosen for the purpose of comparing performance against [1]. However, TabPFN transformer, which is a state-of-the-art neural network, was also additionally used due to its usual superior ability for tabular data classification.

D. Metrics

The metrics used to evaluate these models include the following:

- Accuracy: Accuracy is a metric that describes how the model performs across all classes in general. It is calculated as the ratio between the number of True Positives (TP) and False Positives (FP) to the total number of predictions, that is, the True Negatives (TN), False Negatives (FN), TP and FP.

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN}$$

- F1-score: F1-score is a measure of accuracy that makes use of both the precision and recall metrics in machine learning. A high algorithm F1-score value indicates high accuracy since it considers both false positives and false negatives.

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall}$$

- Sensitivity: Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.

$$Sensitivity = \frac{TP}{TP + FN}$$

- Specificity: Specificity is the metric that evaluates a model's ability to predict true negatives of each available category.

$$Specificity = \frac{TN}{TN + FP}$$

E. Shapley Additive exPlanations (SHAP)

The authors in [9] propose a unified framework for interpreting features importance for a particular prediction. SHAP is based on the Shapley values problem of game theory where the contribution of each member in a coalition C to a coalition value V is measured by averaging the marginal contributions across all possible feature union ordering. SHAP reframes this concept by measuring the importance of each feature x_i to a model f. Equation 1 below shows the Shapley value of a feature 'φi' where z' represents a subset of potential features x', M is the total number of features in the model, and z'\i denotes every potential subset of input features omitting i feature.

$$\varphi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x\left(z'^{\setminus i}\right) \right] \quad (1)$$

F. *Local Interpretable Model-agnostic Explanations (LIME)*

The authors in [11] propose a method to generate an explanation of a prediction by approximating an underlying interpretable local model within a more complex black-box model. This can be achieved by generating a set of perturbed instances and obtaining the predictions using the black-box model. Later, a simple linear model can be applied on the dataset and locally weighted. The local approximation of a black-box model f for an input x is presented in Equation 2 where g is a simple interpretable model that belongs to interpretable models G.

$$\varepsilon(x) = argmin_{g \in G} \mathcal{L}(f, \mathcal{G}, \pi_x) + \Omega(\mathcal{G}) \quad (2)$$

IV. Results

To measure the performance of our classification models, we rely on the accuracy, sensitivity, specificity, and f1-score metrics. Our results show that Light Gradient Boosting (LGB) model outperformers other traditional ML and DL methods with an accuracy of 84.4% and an F1-score of 50.3%. K-Nearest Neighbors (KNN) resulted in an almost similar performance with a slightly lower accuracy and specificity. Nonetheless, all classifiers achieved relatively acceptable performance and illustrated an improvement of the results achieved by [1] on the same dataset. Particularly, LGB and KNN outperform the state-of-the-art results by a factor of almost 10%. The results of this study are summarized in Table 1 with comparison to recently published results in [1] across different models.

TABLE I. Model Performance Summary

Model [1]	**Metrics**			
	Accuracy	***Sensitivity***	***Specificity***	***F1-Score***
SVC	**75.7**	83.7	**74.9**	**40.7**
LR	75.6	83.8	74.6	40.5
KNN	72.9	80.9	72.0	37.3
LGB	73.3	84.3	72.1	38.6
XGB	75.2	83.9	72.0	40.2
CB	74.5	83.6	73.5	39.5
RF	72.5	**84.6**	71.2	38.0
Our method	***Accuracy***	***Sensitivity***	***Specificity***	***F1-Score***
SVC	74.8	86.4	73.6	39.7
LR	66.6	**88.4**	64.3	33.7
KNN	84.3	81.9	84.6	50.1
LGB	**84.4**	81.9	**84.7**	**50.3**
XGB	84.0	82.2	84.2	49.7
CB	83.8	82.8	83.9	49.5
RF	83.7	82.7	83.9	49.4
TabPFN	83.8	82.8	83.9	49.5

To provide a further insight into features contributions to each prediction of the model, SHAP analysis was performed on the best performing model (LGB). Fig. 1. provides the aggregated summary of features effect on model predictions where class 0 denotes 'No Vitamin A deficiency' while class 1 denotes 'Vitamin A deficiency'.

Fig. 1. Global explanations with Shapley Values

To provide explainable artificial intelligence (XAI) predictions, a further insight into the relationship and dependencies between different factors that influences the model prediction was obtained. The dependecy plot of the most important features; Age, Gender, Itching, nightblind, and conjuctivness are shown in Fig. 2, Fig. 3, and Fig. 4.

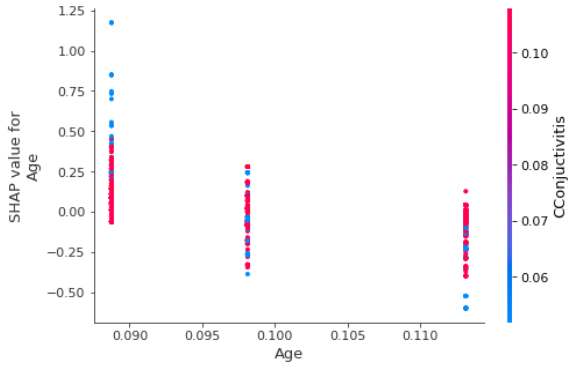

Fig. 2. Dependency Plot of Age with respect to "conjunctivitis"

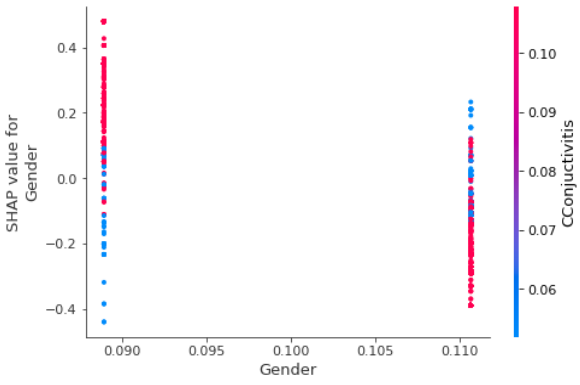

Fig. 3. Dependency Plot of Gender with respect to "conjunctivitis"

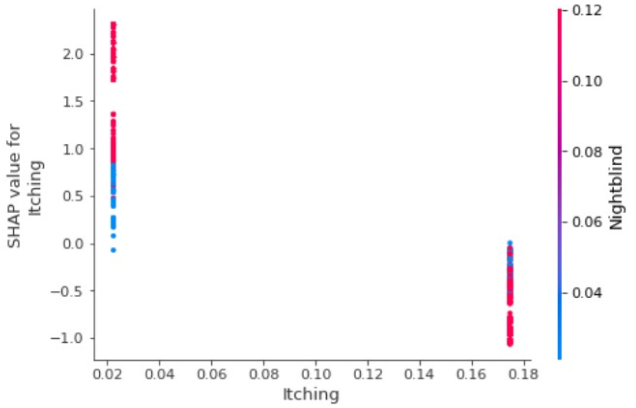

Fig. 4. Dependency Plot of Itching with respect to "Nightblind"

LIME method was later used to obtain further explanation on the model prediction per instance given the probability of the presence of Vitamin A deficiency supported by each feature value and the importance towards the corresponding class. Fig. 5 shows the results of LIME analysis of positive instances of predictions using the best performing model (LGB).

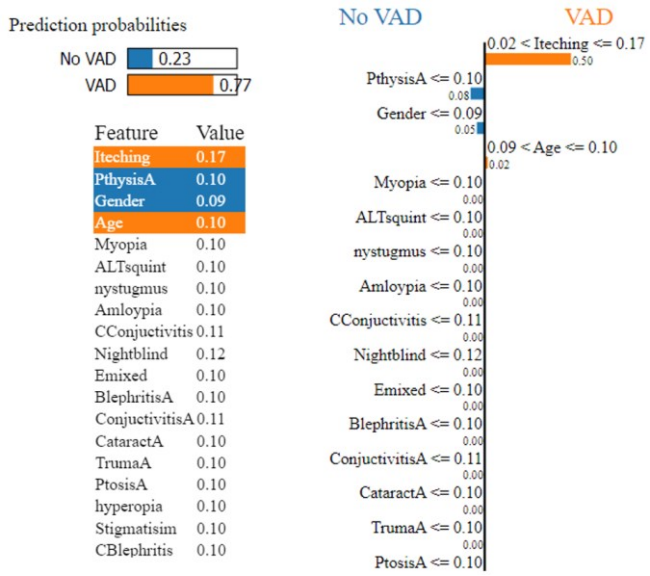

Fig. 5. LIME Analysis results for a predicted presence of Vitamin A Deficiency (VAD) for a positive instance

V. Discussion

Different traditional and ensemble models result in relatively close results with an average accuracy of 80.2% which indicates a form of linearity in the dataset. TabFPN transformer achieved similar results which supports this assumption. Nonetheless, almost all models suffered from an F1-score less than 50% except for LGB and KNN models. This can be a result of the inherited high imbalance within the model features. Unlike previous work [1] which used binary encoding of feature variables, target encoding seems to rapidly improve the performance of ML and DL models. This can be because target encoding replaces the categorical features with a numeric measurement that related the effect of the feature on the target variable which can be helpful for the model and improves its performance. Further, binary encoding can degrade tree-models performance because such models split samples into two bins based on a threshold value with the goal of minimizing impurity of the bins. The split continues based on smaller thresholds as the tree becomes more complicated. Using binary encoded variables can only be split at one point which might negatively affect the overall performance of the model.

The XAI methods used in this study allow us to further investigate the relationship between different features. SHAP analysis illustrated in Fig. 1. shows that gender, age, the presence and absence of itching, and conjunctivitis disorder have the most influence on the models' predictions. Specifically, age and gender seem to have the highest interaction with conjunctivitis disease, where within different levels of diseases night-blind and itching seem to have the maximum interaction On the other hand, features such as alternative squint, nystagmus, phthisis bulbi, amblyopia, exotropia, and esotropia have no influence on the model and can be neglected. The following points further examine the high interactions between age and conjunctivitis, gender and conjunctivitis, and itching and night-blind.

- A deeper look into the relationship and dependencies between feature variables shows that there is a negative linear relationship between the age group and the corresponding SHAP values throughout the three age groups. This indicates that younger age is more associated with the presence of conjunctivitis disorder on average (Fig. 2).

- Gender dependency plot shows that males gender tends to have higher SHAP values related to conjunctivitis in comparison to females (Fig. 3).

- Itching feature dependency plot (Fig. 4) shows a similar trend as the absence of itching tends to have higher SHAP values associated with night-blind feature.

LIME method gives an insight into how much influence each feature has on a certain prediction. Fig. 5 shows that the presence of itching and age group of the patient supported the prediction of having VAD with a probability of 77% with itching having the biggest influence. These XAI approaches would enhance the process of decision making in the field of clinical diagnosis.

VI. Conclusion

In this study, we explored the potential of utilizing Machine Learning and Explainable Artificial Intelligence methods to provide a more-accessible system for Vitamin A deficiency detection using Electrical Health Records of symptoms only. The study enhanced on recently published results in [1] using different pre-processing techniques and discussed two different approaches to explain the relationships between model prediction and the symptoms reported from school children in a low-income country and improve clinical decision making. One possible limitation to this study is the lack of more demographics that can influence the presence of Vitamin A deficiency. Future work should consider including the demographics of the region of interest as it will provide a practically better conclusion.

References

[1] J. Ramesh, S. Donthi, A. Khamis, A. Sagahyroon, and F. Aloul, *Explainable Machine Learning for Vitamin A Deficiency Classification in Schoolchildren*. 2022, p. 04. doi: 10.1109/BHI56158.2022.9926924.

[2] World Health Organization, “Global prevalence of vitamin A deficiency in populations at risk 1995-2005 : WHO global database on vitamin A deficiency,” p. 55, 2009.
[3] “Vitamin A deficiency.” https://www.who.int/data/nutrition/nlis/info/vitamin-a-deficiency (accessed Dec. 13, 2022).
[4] J. N. Nkengasong, K. Yao, and P. Onyebujoh, “Laboratory medicine in low-income and middle-income countries: progress and challenges,” *Lancet*, vol. 391, no. 10133, pp. 1873–1875, May 2018, doi: 10.1016/S0140-6736(18)30308-8.
[5] M. van Lent, W. Fisher, and M. Mancuso, “An Explainable Artificial Intelligence System for Small-unit Tactical Behavior”.
[6] F. Di Martino and F. Delmastro, “Explainable AI for clinical and remote health applications: a survey on tabular and time series data,” *Artif Intell Rev*, Oct. 2022, doi: 10.1007/s10462-022-10304-3.
[7] “A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System - PubMed.” https://pubmed.ncbi.nlm.nih.gov/36298417/ (accessed Dec. 26, 2022).
[8] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[9] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Dec. 13, 2022. [Online]. Available: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
[10] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, “Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features,” *Comput Stat*, vol. 37, no. 5, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.
[11] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” arXiv, Aug. 09, 2016. doi: 10.48550/arXiv.1602.04938.