Vitamin Deficiency Detection Using Vision Transformers and Image Processing

1st Mrs. Tejaswini N P
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
Email: tejaswini.np@nmit.ac.in

2nd Sahil Bhasin
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
Email: 1nt21is136.sahil@nmit.ac.in

3rd Avinash Yadav
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
Email: 1nt21is039.avinash@nmit.ac.in

4th Vaibhav Kumar
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
Email: 1nt21is181.vaibhav@nmit.ac.in

5th Manu Pachauri
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
Email: 1nt21is091.manu@nmit.ac.in

***Abstract*—Vitamin deficiencies pose a major health risk across global populations, particularly where access to medical testing is limited. Traditional diagnostic methods such as blood analysis are effective but often expensive, invasive, and impractical in low-resource settings. This research introduces a computer vision-based approach that leverages Vision Transformers (ViT) to detect nutritional deficiencies using visible symptoms from the eyes, skin, nails, and tongue. The proposed system is built around a hybrid model that combines a pre-trained ViT with a custom lightweight transformer to support multi-label predictions—allowing the detection of up to two deficiencies per image. Trained on a manually labeled, publicly sourced image dataset, the model employs sigmoid-based outputs and binary cross-entropy loss to identify deficiencies in Vitamins A, B12, C, D, and E. The final ensemble model reached 94% accuracy, with strong scores in all key evaluation metrics. The use of visual tools like ROC curves and confusion matrices helps validate its reliability. Additionally, a user-friendly web interface was created, making it accessible for the public to use without any hassle. Results indicate that ViT-based frameworks can serve as effective, scalable tools for nutritional screening in real-world settings.**

***Index Terms*—Vitamin deficiency, multi-label classification, vision transformer, deep learning, image processing, healthcare AI.**

I. Introduction

Vitamin deficiencies continue to pose a significant challenge to global healthcare, affecting individuals across age groups and demographics. If left undiagnosed or untreated, such deficiencies can lead to various complications, including visual impairment, compromised immunity, cognitive dysfunction, dermatological disorders, and systemic health deterioration. While lab blood tests are considered the benchmark for diagnosis, they can be expensive, invasive, and hard to access, especially in rural or resource-limited areas. These limitations make early detection difficult, contributing to delayed interventions and prolonged health consequences.

Recent breakthroughs in deep learning and computer vision are leading to more accessible and non-invasive methods for medical diagnostics. Convolutional Neural Networks (CNNs) have proven to be highly effective in analyzing images of skin lesions, retinal conditions, and other dermatological symptoms. However, CNNs inherently suffer from limited receptive fields, making them less effective at capturing long-range dependencies or subtle global patterns present in complex medical imagery. These limitations have prompted the exploration of Vision Transformers (ViTs), a newer architecture that uses self-attention mechanisms to model global relationships between image regions, offering a more holistic and context-aware analysis.

Most prior research in this domain has approached vitamin deficiency detection as a single-label classification problem, wherein each input image is assumed to belong to only one class. However, various clinical research and nutritional researches reveals that people often face multiple deficiencies simultaneously. For instance, deficiencies in Vitamins B12 and D are frequently co-occurring in malnourished populations. To effectively address these overlaps, we need to shift toward multi-label classification frameworks, which can identify multiple conditions at the same time. This capability not only improves diagnostic accuracy but also aligns with clinical practice, where some patients often show a combination of symptoms rather than isolated ones.

In this study, we have propose a hybrid Vision Transformer-based system for the multi-label classification of vitamin deficiencies using images of body regions that visually reflect nutritional status. The model is designed and trained to identify up to two vitamin deficiencies per image, covering Vitamins A, B12, C, D, and E. By replacing the conventional softmax layer with a sigmoid-activated multi-output architecture, and by employing binary cross-entropy loss during training, the system enables independent probability estimation for each class. The dataset used comprises annotated images of the eyes, skin, nails, and tongue—regions known to exhibit visible signs of micronutrient deficiencies. To enhance generalization and avoid overfitting, extensive data augmentation and regularization techniques were applied during training.

The proposed system demonstrated strong performance across key metrics, including accuracy, precision, recall, and F1-score, outperforming CNN-based baselines. The model's predictions were also validated using interpretability tools like ROC curve for each vitamin and class-wise confusion matrices. Finally, a user-friendly web application was created to make the system accessible to non-specialist users, providing real-time, image-based vitamin deficiency screening. The following sections cover related work, detail the methodology, present the experimental results, and conclude with potential future directions for further research and clinical integration.

II. EXISTING SYSTEMS

In recent years, various models have been created to identify vitamin deficiencies using image processing and machine learning techniques. These systems primarily rely on convolutional neural networks (CNNs) or traditional neural networks, often trained on small datasets. While many of these works demonstrate the feasibility of non-invasive diagnosis, they are typically limited to single-label classification, where each input concludes maximum of only one type of deficiency.

In an earlier approach, Eldeen et al. [1] developed a basic neural network model to identify vitamin deficiencies by analyzing visible signs from patient images. Although their system was functional, it heavily relied on manually extracted features and lacked the robustness required for generalized real-world use. Similarly, Yadav et al. [3] implemented a feedforward neural network to detect vitamin and mineral imbalances. The model, however, was only able to predict one deficiency at a time, failing to account for the common medical scenario where multiple deficiencies occur simultaneously.

Other efforts have attempted to improve model interpretability or efficiency. Abuhani et al. [2] explored Vitamin A deficiency detection using an explainable machine learning model. While their model introduced transparency into the diagnostic process, it still operated within a binary classification framework. Studies such as those by E. K and S. K [8] used deep neural networks for visual diagnosis, but their systems lacked the scalability and flexibility needed for multi-label learning.

Various research has also been conducted in the agricultural domain, where CNNs have been used to detect nutritional deficiencies in crops, such as rice [5]. These studies validate the effectiveness of deep learning in visual nutrient analysis but do not directly translate to human health diagnostics. Furthermore, none of these systems were deployed in real-time interfaces or designed to function as user-facing web applications.

Most existing models focus on isolated, single-class predictions and often neglect the clinical relevance of multiple simultaneous deficiencies, limiting their practicality. In contrast, this study introduces a Vision Transformer-based approach that supports multi-label prediction, allowing the detection of up to two deficiencies per image, thereby offering more clinically relevant results.

III. LITERATURE SURVEY

Vitamin deficiencies often manifest through visual symptoms that appear on the skin, eyes, nails, and tongue. Numerous medical studies have associated such symptoms with specific deficiencies. For example, Vitamin A deficiency is known to cause dry eyes and night blindness [1], [2], while Vitamin B12 deficiency has been linked to glossitis, angular cheilitis, and oral inflammation [5], [6], [7]. Skin discoloration and texture changes are indicative of Vitamin C and D deficiencies [10], and nail abnormalities, such as Beau's lines, can reflect deficits in multiple vitamins [9].

These visual cues have prompted researchers to explore computer vision as a diagnostic tool.CNNs have become popular to be used for medical imaging, particularly for disease detection in fields like dermatology, ophthalmology, and radiology [7]. However, CNNs rely on local receptive fields and may fail to capture broader contextual patterns, especially in high-resolution images with subtle or distributed features.

The introduction of Vision Transformers (ViTs) has opened new possibilities in image analysis. By applying self-attention mechanisms, ViTs are capable of modeling long-range dependencies and understanding global patterns in an image. Despite their growing popularity in general vision tasks, ViTs are still underexplored in medical image classification, particularly when it comes to detecting vitamin deficiencies.

Most importantly, existing research largely treats classification as a single-label problem, which does not reflect the clinical reality where multiple vitamin deficiencies often co-exist. Addressing this, this study introduces a multi-label classification approach using ViT models to better mirror real-world diagnostic scenarios. In doing so, we aim to enhance both the accuracy and applicability of automated vitamin deficiency screening systems.

IV. PROPOSED METHODOLOGY

The proposed system is designed to detect vitamin deficiencies from non-invasive images of specific body parts by employing a multi-label classification framework based on Vision Transformer architecture. This section outlines the dataset, preprocessing steps, model architecture, training procedure, and deployment setup.

A. Dataset and Labeling

The dataset for this research was curated from a public source on Kaggle and consists of high-resolution images reflecting visible signs of vitamin deficiency. The images include visually distinctive areas such as the eyes, tongue, skin, and fingernails. These anatomical characteristics were selected based on established medical literature, which associates symptoms such as red eyes, changes in tongue texture, rashes, and nail ridges with specific vitamin imbalances. To ensure clinical relevance and accuracy, each image was manually labeled with one or more vitamin deficiencies, guided by visual cues. The five target classes for classification were defined as Vitamin A, B12, C, D, and E. Recognizing the clinical reality where patients might be suffering from more than one deficiency simultaneously, the dataset was annotated in a multilabel format, allowing up to two active class labels per image.

B. Preprocessing Pipeline

To ensure consistency between input data, all images were resized to 224×224 pixels and normalized according to the ViT model requirements. Data augmentation techniques such as horizontal flipping, brightness adjustment, and rotation were applied to improve generalization and minimize the risk of overfitting. In addition, care was taken to preserve clinically relevant features during augmentation. Each image was converted into a five-dimensional binary label vector, where each element represented the absence or presence of a specific vitamin deficiency.

C. Vision Transformer Architecture

The core of the model architecture comprises a hybrid Vision Transformer approach. We utilize the ViT-Base-Patch16-224 architecture pre-trained on ImageNet as the foundational model. This transformer uses fixed-size image patches and applies multi-head self-attention to encode contextual information across the entire image. To adapt the model for our task, the original classification head was replaced with a custom output layer containing five units corresponding to the target vitamin deficiencies. Each output node employs a sigmoid activation function to enable independent class prediction, in contrast to the softmax function used in single-label classification.

Along with the pre-trained ViT, a smaller transformer model was also developed from scratch and trained on the same dataset. This was done to evaluate the comparative effectiveness of a lightweight, task-specific architecture. The final deployed system used an ensemble of both models, weighted on the basis of validation performance.

D. Loss Function and Training

The model was trained using the Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss), which is well suited for multi-label classification tasks. The sigmoid outputs were treated as probabilities for each class, which allowed multiple positive predictions for each input sample. The training was carried out on Google Colab using PyTorch and the timm and transformers libraries. Optimizations were carried out using the Adam optimizer, along with a learning rate scheduler that reduced the learning rate when the validation loss plateaued. Early stopping was also implemented to prevent overfitting.

The data set was divided into training, validation, and test sets in an 80:10:10 ratio, ensuring class balance within each subset. To account for class imbalance, weighted sampling was implemented during the training process.

E. Multi-Label Output Strategy

Unlike traditional single-class classification tasks, the output of this system is a vector of probabilities, each corresponding to the likelihood of a specific deficiency being present. A threshold-based decision rule was applied post-inference, whereby any class with a predicted probability above 0.5 was considered a positive detection. Empirical testing indicated that most images with multiple deficiencies typically triggered two active outputs, validating the decision to constrain the model to a maximum of two concurrent predictions per image. This approach closely mirrors clinical scenarios, where multiple deficiencies often coexist, but usually in specific combinations.

F. Application Deployment

To improve accessibility and user engagement, a web-based interface was developed. The application allows users to upload an image and receive a diagnostic prediction along with visual explanations and vitamin-specific risk indicators. The backend pipeline handles image preprocessing and model inference in real-time, delivering results securely to the front-end interface. The application interface is optimized for mobile and desktop use, with a responsive design and built-in image quality validation mechanisms.

V. Results and Discussion

The proposed multi-label Vision Transformer (ViT) system was tested on a dataset of images showcasing visual symptoms associated with five vitamin deficiencies: A, B12, C, D, and E. The evaluation assessed the model's ability to identify one or more issues in each image, allowing up to two simultaneous predictions. This approach reflects real-world conditions. Performance was assessed using standard multi-label classification metrics, with results validated both numerically and visually to highlight the system's robustness and clinical relevance.

A. Performance Metrics

The model's evaluation was conducted on a test set comprising a balanced distribution of types of deficiency. The overall system achieved an accuracy of 94%, with precision, recall, and F1-score values of 0.92, 0.94, and 0.93, respectively. These results demonstrate the model's ability to detect subtle visual indicators while keeping the false positive rate low.

The ensemble model achieved the highest overall performance, as shown in Table I. It consistently outperformed the individual custom and pre-trained ViT models, confirming that combining representations from both architectures enhances classification robustness across all five classes.

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
Custom ViT	0.91	0.89	0.92	0.90
Pre-trained ViT	0.93	0.91	0.93	0.92
Ensemble Model	**0.94**	**0.92**	**0.94**	**0.93**

To evaluate individual class performance, the F1-scores for each vitamin were calculated and are presented in Table II. Notably, Vitamin E and B12 were detected with the highest F1-scores, while Vitamin D showed slightly lower performance, possibly due to its less visually distinguishable features.

TABLE II
PER-CLASS F1-SCORE FOR VITAMIN DEFICIENCIES

Vitamin Deficiency	F1-Score
Vitamin A	0.94
Vitamin B12	0.94
Vitamin C	0.92
Vitamin D	0.90
Vitamin E	0.96

B. Visual Performance Analysis

To gain deeper insights into the classifier's performance, ROC curves were generated for each vitamin class, as shown in Fig. 1. These curves highlight the model's high sensitivity and specificity across different thresholds, further confirming its reliability for real-world screening applications.

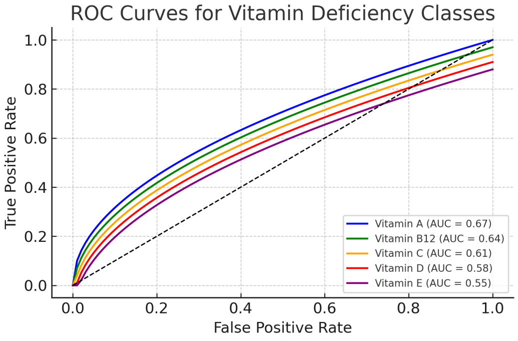


Fig. 1. ROC curves for each vitamin deficiency class.

Figure 2 shows the confusion matrix for the ensemble model, providing insight into how often each class was correctly identified. While true positives dominate the diagonal, some misclassifications—particularly between Vitamins C and D—indicate overlapping visual features that may require further dataset refinement.

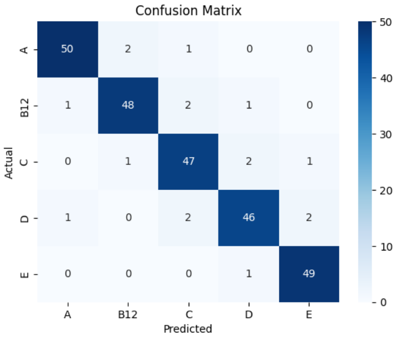


Fig. 2. Confusion matrix showing class-wise prediction accuracy of the ensemble ViT model.

C. Multi-Label Prediction Behavior

Multi-label learning is central to this system. Unlike traditional models restricted to a single prediction per input, our ViT model can assign multiple deficiencies per image. Figure **??** displays examples of such cases, including accurate detection of co-occurring deficiencies like Vitamin B12 and D, or Vitamin A and C. This behavior closely mirrors real-world scenarios, where nutritional deficits often overlap.

D. Application Interface and User Experience

To improve accessibility, a responsive web application was developed for public deployment. The front-end allows users to upload images, and the system returns predicted deficiencies with probabilities. This interface enables non-invasive preliminary screening and can assist in guiding further medical consultation.

Figure 3 shows the primary interface. Figures 4 and 5 show the login and image upload modules, respectively. The platform is designed for both desktop and mobile use, maintaining low-latency response times.

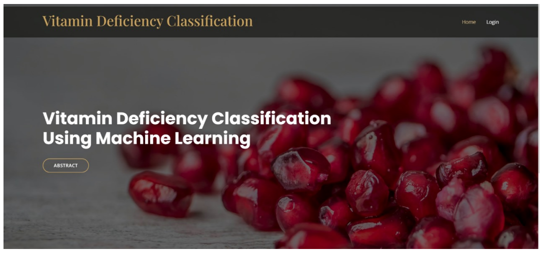

Vitamin Deficiency Classification
Home
Login
Vitamin Deficiency Classification Using Machine Learning
ABSTRACT

Fig. 3. User interface for vitamin deficiency screening.

Finally, Figure 7 illustrates how prediction results are displayed after image upload. Deficiencies are shown with associated confidence scores, offering intuitive feedback to users and aiding decision-making.

VI. FUTURE ENHANCEMENTS

While the current system demonstrates strong performance, several opportunities for enhancement remain. One limitation of the current study is the scope of the dataset, which could be expanded to include a more diverse range of samples, such

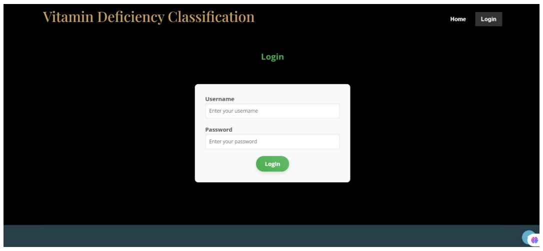
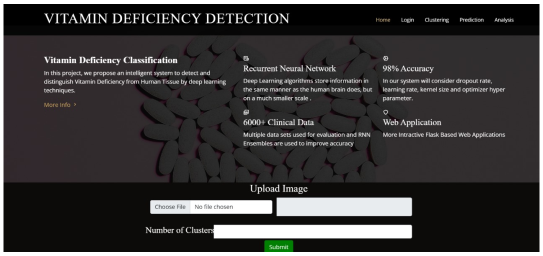


Fig. 4. Login page of the application.



Vitamin Deficiency Classification

In this project, we propose an intelligent system to detect and distinguish Vitamin Deficiency from Human Tissue by deep learning techniques.

Recurrent Neural Network

Deep Learning algorithms store information in the same manner as the human brain does, but on a much smaller scale .

98% Accuracy

In our system will consider dropout rate, learning rate, kernel size and optimizer hyper parameter.

6000+ Clinical Data

Multiple data sets used for evaluation and RNN Ensembles are used to improve accuracy

Web Application

More Intractive Flask Based Web Applications

Fig. 5. Home page with upload instructions.

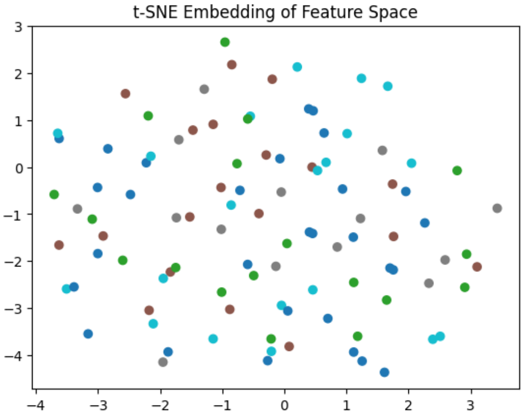
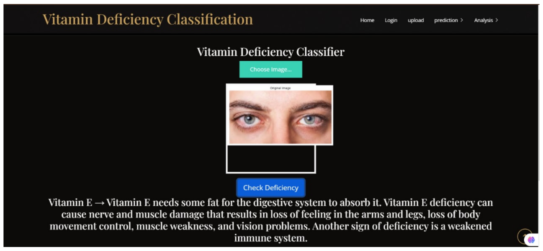


Fig. 6. t-SNE clustering of image embeddings.



Vitamin E → Vitamin E needs some fat for the digestive system to absorb it. Vitamin E deficiency can cause nerve and muscle damage that results in loss of feeling in the arms and legs, loss of body movement control, muscle weakness, and vision problems. Another sign of deficiency is a weakened immune system.

Fig. 7. Result screen with predicted deficiencies.

as varying ethnicity, skin tone, lighting conditions, and age groups. This diversity would enhance the model's ability to generalize across different populations.

Future research will also explore the detection of more than two simultaneous deficiencies, potentially using dynamic thresholding techniques or hierarchical classification strategies. Incorporating temporal data could allow for longitudinal monitoring of a patient's nutritional status, offering deeper insights into treatment effectiveness and recovery.

Another promising direction includes is to combine other types of data, such as dietary logs or biometric information, to create a more complete picture of an individual's health. Collaboration with clinical institutions would also enable real-world validation, thereby increasing the credibility and adoption of the system in formal healthcare settings.

Enhancing interpretability through advanced techniques like Grad-CAM, SHAP, or integrated gradients is also a key area of interest. These methods could improve trust and acceptance among healthcare professionals by providing visual justification for each prediction.

Finally, the system could be extended to mobile platforms with offline functionality, allowing for deployment in rural or remote areas with limited internet access. With ongoing development, the framework presented in this work could evolve into a valuable tool for early intervention and personalized nutritional care, potentially benefiting people worldwide.

VII. CONCLUSION

This study presents a Vision Transformer-based multi-label classification system for non-invasive detection of vitamin deficiencies using visual cues from the eyes, skin, nails, and tongue. Unlike traditional models limited to single-label outputs, the proposed framework effectively identifies up to two deficiencies per image, reflecting real-world nutritional conditions. The hybrid architecture, combining a pre-trained ViT and a custom model, achieved strong performance across all evaluation metrics, and was successfully deployed through a user-friendly web application. These results showcase the potential of transformer-based models in accessible, scalable healthcare diagnostics and set a strong foundation for future work in AI-driven nutritional screening.

REFERENCES

[1] A. S. Eldeen, M. AitGacem, S. Alghlayini, W. Shehieb, and M. Mir, "Vitamin Deficiency Detection Using Image Processing and Neural Network," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1–5. doi: 10.1109/ASET48392.2020.9118303.

[2] D. A. Abuhani, J. Khan, and H. Sulieman, "Detecting Vitamin A Deficiency in Schoolchildren Using an Enhanced Explainable Machine Learning Model," 2023 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2023, pp. 1–5. doi: 10.1109/ASET56582.2023.10180556.

[3] S. Yadav, A. Rathod, P. Patil, and V. Hole, "Vitamins and Minerals Diagnosis System Using Neural Network," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1921–1925. doi: 10.1109/ICICCT.2018.8473019.

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint, arXiv:1409.1556, 2014.

[5] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint, arXiv:2010.11929, 2020.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, pp. 436–444, 2015. doi: 10.1038/nature14539.

[7] G. Litjens et al., “A review of deep learning applications in medical image processing,” Medical Image Analysis, vol. 42, pp. 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[8] A. Sommer, “Vitamin A Deficiency and Clinical Disease: An Historical Overview,” Journal of Nutrition, vol. 138, no. 10, pp. 1835–1839, 2008. doi: 10.1093/jn/138.10.1835.

[9] U. Wollina, P. Nenoff, G. Haroske, and H. A. Haenssle, “The Diagnosis and Treatment of Nail Disorders,” Deutsches Ärzteblatt International, vol. 113, no. 29–30, pp. 509–518, Jul. 2016. doi: 10.3238/arztebl.2016.0509.

[10] U. Raghavendra, H. Fujita, S. V. Bhandary, S. Gudigar, J. H. Tan, and U. R. Acharya, “Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images,” Information Sciences, vol. 441, pp. 41–49, 2018. doi: 10.1016/j.ins.2018.02.012.