JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Hybrid CNN–Vision Transformer Framework with Swin Transformer for Image-Based Vitamin Deficiency Detection

[1]A.Vanaja, [2] I.Kullayamma

[1]M.Tech Student, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

[2]Professor, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

Abstract : **Vitamin deficiencies can produce visible changes in body regions such as the skin, eyes, lips, nails, and tongue, providing an opportunity for image-based preliminary screening. This paper proposes a deep-learning-based Vitamin Deficiency Detection System using an Enhanced Convolutional Neural Network (CNN), Vision Transformer (ViT), and Swin Transformer with attention mechanisms. The Enhanced CNN, based on EfficientNet-B3, extracts local spatial and texture features while spatial attention emphasizes informative regions. The Vision Transformer captures global relationships among image patches, whereas the Swin Transformer learns hierarchical representations using shifted-window attention. The prediction scores from the three models are combined through an ensemble score-fusion mechanism and evaluated using a threshold-based decision process. The system classifies the input as healthy or deficiency-related and provides the corresponding vitamin category and severity information. Experimental evaluation demonstrates model scores of 49.0%, 47.5%, and 74.1% average for CNN, ViT, and Swin Transformer, respectively, demonstrating the potential of the proposed hybrid framework for preliminary image-based vitamin deficiency screening.**

Index Terms **- Vitamin Deficiency Detection, Deep Learning, Enhanced CNN, EfficientNet-B3, Vision Transformer (ViT), etc.**

I. INTRODUCTION

Vitamin deficiencies are common nutritional disorders that can lead to various health complications, making early identification important for timely intervention. Conventional assessment generally relies on clinical examination and laboratory-based investigations. However, the availability of visible manifestations associated with some deficiencies has created an opportunity for developing non-invasive, image-based preliminary screening approaches. Changes in regions such as the skin, eyes, lips, tongue, and nails can provide visual information that may be analyzed computationally. Eldeen et al. demonstrated the feasibility of detecting possible vitamin deficiencies from images of human body regions using image processing and neural networks, establishing an important foundation for image-based vitamin deficiency assessment [4]. Machine learning and deep learning have subsequently been explored for automated vitamin deficiency analysis. Sambasivam et al. investigated machine-learning techniques for predicting Vitamin D deficiency severity using clinical and lifestyle-related parameters [1]. Elavarasi and Shanmugapriya developed a Deep Neural Network (DNN)-based approach for vitamin deficiency diagnosis [2]. Image-processing-based neural-network approaches have also been investigated for identifying deficiency-related visual characteristics [4], [15]. More recently, Bonde et al. employed a Deep-CNN-based approach to analyze images of regions such as the skin, eyes, and nails [16]. Jayaram et al. compared DenseNet, ResNet, CNN, and VGG16 for image-based vitamin deficiency prediction and reported the effectiveness of DenseNet for the considered task [17]. Despite these developments, several challenges remain. First, visible deficiency symptoms can be subtle and may differ significantly among individuals. Second, image acquisition conditions such as illumination, brightness, contrast, camera quality, and skin appearance can affect the visual characteristics extracted from an image. Third, conventional image-processing methods generally depend on manually designed color, texture, and statistical features, which may not sufficiently represent complex deficiency-related patterns. CNN-based approaches can automatically learn local spatial and texture features, but their ability to model long-range relationships across the complete image can be limited. The need to capture complementary local and global information is therefore an important challenge in image-based deficiency detection [3], [5]–[14]. Recent medical image-analysis research has demonstrated the benefits of combining different deep-learning architectures. Parallel CNNs and ensemble models have been investigated for medical image classification [6], [7], while hybrid architectures have been used to combine complementary feature representations [10], [14]. These studies indicate that combining multiple models can provide richer representations than relying on a single architecture. However, existing vitamin deficiency studies have mainly focused on conventional image processing, DNNs, CNNs, or individual pretrained architectures [2], [4], [15]–[17]. Limited research has investigated the combination of CNN-based local features with Transformer-based global and hierarchical representations for human vitamin deficiency screening.

The motivation of this work is to develop a robust image-based framework that can analyze complementary visual characteristics from different human body regions. To address the limitations of individual models, the proposed work integrates an Enhanced Convolutional Neural Network (CNN), Vision Transformer (ViT), and Swin Transformer. The Enhanced CNN, based on EfficientNet-B3, is used to extract local spatial and texture features, while spatial attention emphasizes informative image regions. ViT is employed to model global relationships between image patches, whereas Swin Transformer provides

hierarchical feature representations through shifted-window attention. The outputs of these models are combined using an ensemble score-fusion mechanism.

The main objectives of this work are:

- To develop an image-based framework for preliminary vitamin deficiency screening.
- To extract local spatial and texture features using an Enhanced EfficientNet-B3-based CNN.
- To employ attention mechanisms for emphasizing informative regions in input images.
- To utilize ViT for capturing global relationships among image patches.
- To utilize Swin Transformer for hierarchical feature extraction.
- To combine the outputs of CNN, ViT, and Swin Transformer using ensemble score fusion.
- To classify images as healthy or deficiency-related and provide corresponding vitamin and severity information.

The major contributions of this work are:

- An Enhanced CNN–ViT–Swin Transformer ensemble framework is proposed for image-based vitamin deficiency screening.
- EfficientNet-B3 is enhanced with a spatial attention mechanism for local feature extraction.
- ViT is incorporated to capture global relationships within the input image.
- Swin Transformer is employed to learn hierarchical visual representations.
- A multi-model ensemble mechanism is used to combine the complementary prediction scores.
- The proposed framework considers visible regions including the skin, eyes, lips, nails, and tongue for preliminary deficiency analysis.
- The system provides vitamin-category and severity-related information along with the classification result.

The remainder of this paper is organized as follows. Section II presents the literature survey and discusses existing approaches related to vitamin deficiency detection and deep-learning-based medical image analysis. Section III describes the proposed methodology, including image preprocessing, Enhanced CNN, spatial attention, ViT, Swin Transformer, and ensemble score fusion. Section IV presents the experimental setup and simulation results. Section V discusses the obtained results and their implications. Finally, Section VI concludes the paper and presents the future scope of the proposed work

II. LITERATURE SURVEY

Sambasivam et al. [1] investigated the prediction of Vitamin D deficiency severity using machine learning techniques. The study analyzed clinical, demographic, and lifestyle-related parameters and evaluated multiple machine learning algorithms for deficiency severity prediction. The work demonstrates the applicability of machine learning for automated vitamin deficiency assessment; however, it primarily relies on non-image-based parameters. E. K. and S. K. [2] proposed a Deep Neural Network (DNN)-based approach for diagnosing vitamin deficiencies in humans, demonstrating the potential of deep learning for automated deficiency identification. Gul and Bora [3] investigated pretrained Convolutional Neural Networks (CNNs) for detecting nutrient deficiencies in hydroponic basil. Their work demonstrated that pretrained CNN architectures can effectively learn visual characteristics associated with deficiency symptoms. Although the application is related to plant nutrient deficiency, the study provides motivation for using transfer-learning-based visual feature extraction. Eldeen et al. [4] proposed an image-processing and neural-network-based approach for vitamin deficiency detection using visible symptoms from human body regions. The study established the feasibility of utilizing images for preliminary vitamin deficiency assessment. Solanki et al. [5] presented a comprehensive study of intelligent techniques for brain tumor detection and classification, highlighting the increasing application of deep learning in medical image analysis. Rahman and Islam [6] developed a parallel deep CNN framework for brain tumor detection and classification from MRI images. Patil and Kirange [7] proposed an ensemble of deep learning models for brain tumor detection, demonstrating the potential of combining multiple deep-learning models for improved classification. Satyanarayana et al. [8] developed a deep CNN-based approach incorporating mass-correlation-based feature analysis for brain tumor classification. Gupta et al. [9] investigated a deep CNN-based brain tumor detection framework, demonstrating the effectiveness of automatically learned image features. Akter et al. [10] proposed a CNN and U-Net-based framework for MRI classification and tumor segmentation, illustrating the advantages of combining complementary deep-learning architectures. Kumar and Kumar [11] investigated CNN-based brain tumor classification and segmentation, further demonstrating the applicability of deep convolutional features for medical image analysis. Mohsen et al. [12] combined a single-image super-resolution technique with pretrained ResNeXt101 and VGG19 models for brain tumor classification, showing the benefit of image enhancement and pretrained architectures. Rajendran et al. [13] investigated automated brain tumor MRI segmentation using deep learning, emphasizing automated feature learning for medical image analysis. Jabbar et al. [14] proposed a hybrid Caps-VGGNet model for brain tumor detection and multi-grade segmentation, demonstrating the potential of hybrid deep-learning architectures. Hipparkar [15] investigated vitamin deficiency detection using image processing and neural networks. The work focused on visible symptoms associated with vitamin deficiencies and demonstrated the feasibility of image-based deficiency identification. Bonde et al. [16] proposed a vitamin deficiency detection system using image processing and a Deep-CNN algorithm, highlighting the capability of CNN-based feature extraction for analyzing visual symptoms. Jayaram et al. [17] investigated a DenseNet-based deep-learning technique for vitamin deficiency prediction and compared deep architectures for visual deficiency classification. From the above studies, it can be observed that existing vitamin deficiency detection approaches mainly employ machine learning, image processing, DNNs, CNNs, or individual pretrained deep-learning architectures [1]–[4], [15]–[17]. The medical image analysis studies further demonstrate the effectiveness of ensemble, hybrid, and multi-architecture deep-learning approaches [5]–[14]. However, limited work has explored the combination of CNN-based local feature extraction with Vision Transformer-based global representation and Swin Transformer-based hierarchical feature learning for human vitamin deficiency detection. Therefore, this work proposes an Enhanced CNN–ViT–Swin Transformer ensemble framework with attention mechanisms for image-based

preliminary vitamin deficiency screening. The proposed framework combines complementary local, global, and hierarchical image representations to improve the analysis of visible deficiency-related characteristics.

Table 1: Literature Summery

Ref.	Main Technique	Application	Key Contribution
[1]	ML, Random Forest	Vitamin D deficiency	Severity prediction using clinical/lifestyle parameters
[2]	DNN	Human vitamin deficiency	Deep-learning-based deficiency diagnosis
[3]	Pretrained CNNs	Plant nutrient deficiency	Transfer learning and Grad-CAM
[4]	Image Processing + NN	Human vitamin deficiency	Eye, lip, tongue and nail image analysis
[5]	Intelligent/DL methods	Brain tumor	Review of intelligent medical-image techniques
[6]	Parallel CNN	Brain tumor	Multiple CNN-based feature learning
[7]	Ensemble DL	Brain tumor	Feature fusion and ensemble classification
[8]	CNN + mass correlation	Brain tumor	Feature weighting and preprocessing
[9]	Deep CNN	Brain tumor	Automated deep feature extraction
[10]	CNN + U-Net	Brain tumor	Classification and segmentation
[11]	CNN	Brain tumor	Classification and segmentation
[12]	ResNeXt + VGG19	Brain tumor	Super-resolution and pretrained models
[13]	Deep Learning	Brain tumor	Automated MRI segmentation
[14]	CapsNet + VGGNet	Brain tumor	Hybrid spatial-feature representation
[15]	Image Processing + NN	Vitamin deficiency	Human body-region image analysis
[16]	Deep CNN	Vitamin deficiency	CNN-based visual classification
[17]	DenseNet	Vitamin deficiency	Deep feature extraction and improved classification

III. Proposed Method

A hybrid deep-learning framework for image-based preliminary vitamin deficiency screening by integrating an Enhanced Convolutional Neural Network (CNN), Vision Transformer (ViT), and Swin Transformer. Initially, the input images of the skin, eyes, lips, nails, and tongue are validated and preprocessed through resizing, brightness and contrast enhancement, CLAHE-based enhancement, and normalization. The Enhanced CNN, implemented using an EfficientNet-B3 backbone with spatial attention, extracts discriminative local spatial, texture, and appearance features from the input image. Subsequently, the Vision Transformer divides the image into fixed-size patches and employs self-attention to capture global relationships among different image regions. The Swin Transformer further extracts hierarchical representations using shifted-window attention, enabling the model to learn both local and cross-window visual patterns. The prediction scores generated by the three models are combined using an ensemble score-fusion mechanism and evaluated against a predefined threshold. Based on the fused score, the system classifies the input as healthy or deficiency-related and provides the corresponding vitamin category and severity information. This combination of complementary CNN and Transformer representations is designed to improve the robustness of image-based vitamin deficiency screening.

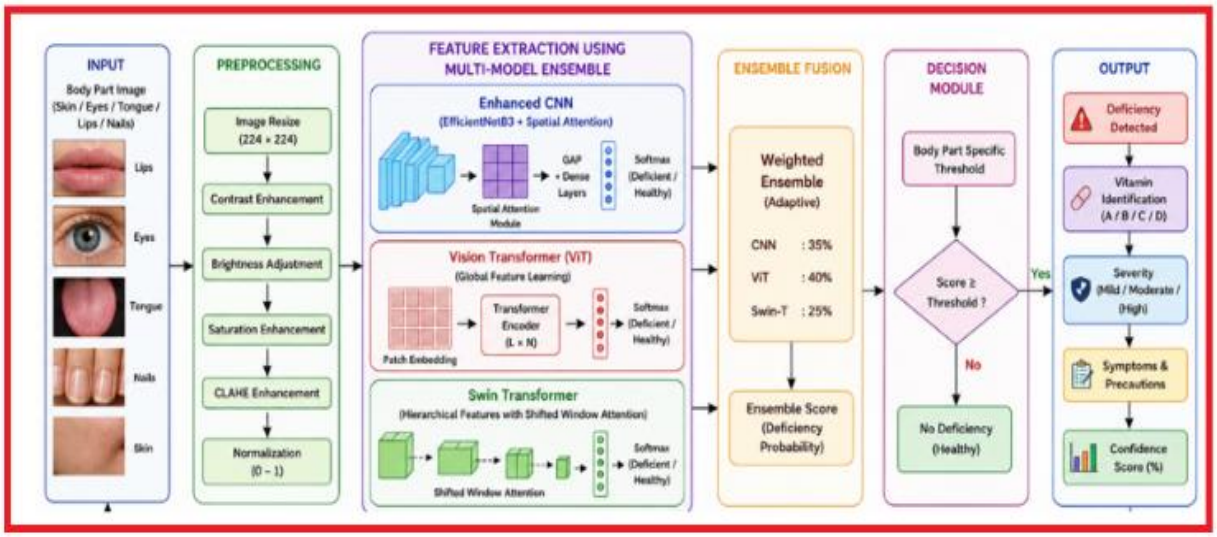

Fig.1: Proposed Architecture of the Hybrid CNN–ViT–Swin Transformer-Based Vitamin Deficiency Detection System

Fig.1 shows that the proposed architecture consists of five major stages: input, preprocessing, multi-model feature extraction, ensemble fusion, and decision/output generation, as shown in the proposed architecture. The system accepts images of different body regions, including the skin, eyes, lips, tongue, and nails, as input. In the preprocessing stage, the input image is resized to 224 × 224 pixels and enhanced using contrast, brightness, saturation, and CLAHE-based enhancement, followed by normalization. The pre-processed image is then processed through three complementary deep-learning models. The Enhanced CNN based on EfficientNet-B3 with spatial attention extracts local spatial and texture features, while the Vision Transformer (ViT) captures global

relationships through patch embedding and self-attention. The Swin Transformer extracts hierarchical visual features using shifted-window attention. Each model produces an individual prediction score. These scores are combined using weighted ensemble fusion, with the architecture indicating weights of CNN = 35%, ViT = 25%, and Swin-T = 40%. The resulting ensemble score is passed to the decision module, where it is compared with a body-part-specific threshold. If the score exceeds the threshold, the system identifies a deficiency and provides the corresponding vitamin category, severity, symptoms, precautions, and food information. Otherwise, the input is classified as having no deficiency, and the final classification score is displayed

3.1 Methodology

Integrates an Enhanced CNN, Vision Transformer (ViT), and Swin Transformer to extract complementary visual features from images of the skin, eyes, lips, tongue, and nails. The overall process consists of image preprocessing, multi-model feature extraction, weighted ensemble fusion, and threshold-based decision making.

Input Dataset

The first stage of the proposed system is image acquisition. The user provides an image of a selected body region through the system interface.

Fig.2: Input Dataset

Table 2 presents the body-region-wise association used in the proposed system for preliminary vitamin deficiency screening. The framework considers skin and eyes for Vitamin A, tongue for Vitamin B, lips for Vitamin C, and nails for Vitamin D. These associations are used by the decision module to relate the detected visual characteristics to the corresponding vitamin category and generate the final screening output.

Table 2: Literature Summery Mapping of Body Regions to Associated Vitamin Categories

Body Region	Associated Vitamin Category
Skin	Vitamin A
Eyes	Vitamin A
Tongue	Vitamin B
Lips	Vitamin C
Nails	Vitamin D

Image Preprocessing

The input image $I$is first resized to 224×224pixels to provide a uniform input to the deep-learning models:

$$I_r = \text{Resize}(I, 224 \times 224) \quad (1)$$

Brightness, contrast, and saturation enhancement are subsequently applied, followed by CLAHE enhancement to improve local contrast. The resulting image is normalized as

$$I_n = \frac{I_r - \mu}{\sigma} \quad (2)$$

where $\mu$and σrepresent the mean and standard deviation used for normalization.

+

An Enhanced CNN based on EfficientNet-B3 is employed to extract local spatial, texture, and appearance features. Spatial attention is incorporated to emphasize informative regions. For an intermediate feature map F, the attention map can be represented as

$$A = \sigma\left(\text{Conv}_{1\times1}(F)\right) \quad (3)$$

where σdenotes the sigmoid activation. The attention-refined feature map is obtained as

$$F_A = F \odot A \quad (4)$$

where ⊙ represents element-wise multiplication. Global average pooling is then applied to obtain the CNN representation, which is passed to the classification layer to generate the CNN score S_{CNN}. Fig.3 shows that the Enhanced CNN architecture uses a pretrained EfficientNet-B3 backbone to extract hierarchical features from the pre-processed 224 X 224input image. The extracted feature map is passed to a Spatial Attention Module, where a 1X1 convolution and sigmoid activation generate an attention map to emphasize informative regions. The refined features are then processed using Global Average Pooling to obtain a compact feature representation. Finally, a fully connected classification head with batch normalization, ReLU activation, and dropout generates the deficiency probability scores for the different vitamin classes. This architecture enables the CNN to focus on relevant visual characteristics while reducing the influence of less informative regions

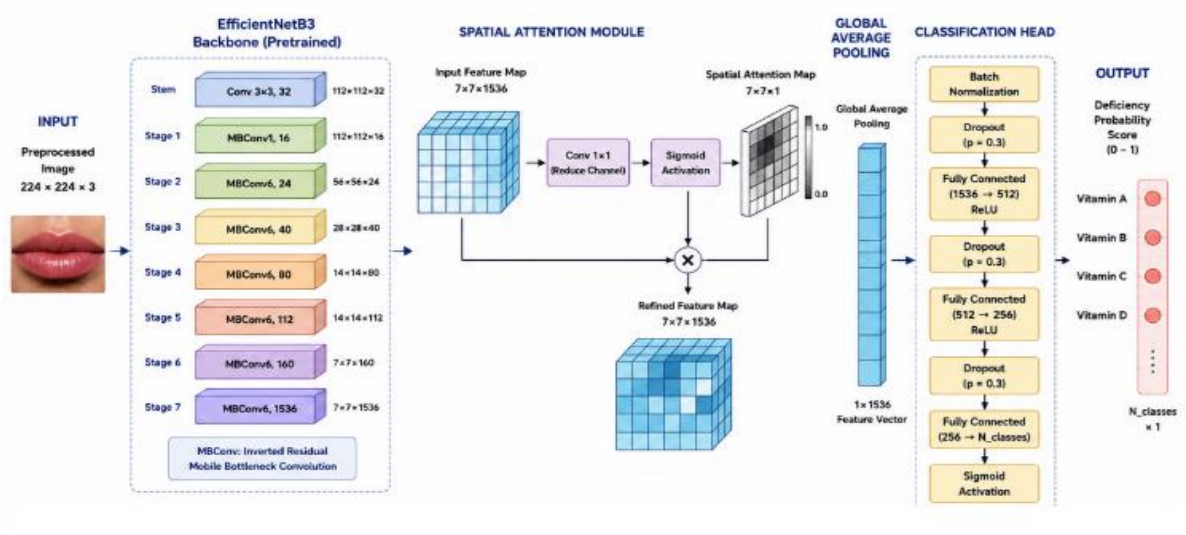

Fig.3: Enhanced CNN Architecture Network

Fig.4 shows the EfficientNet-B3 pretrained backbone used in the proposed Enhanced CNN. The 224x224 input image is passed through the stem and multiple MBConv stages, where depthwise convolution and Squeeze-and-Excitation (SE) blocks progressively extract hierarchical spatial and texture features. The network reduces the spatial dimensions while increasing the number of feature channels, producing a 7x7x1536 feature representation at the final stage. The MBConv block uses expansion convolution, depthwise 3x3 convolution, SE-based channel attention, and projection convolution, with residual connections where applicable. This backbone provides the deep feature representation that is subsequently supplied to the spatial attention and classification modules.

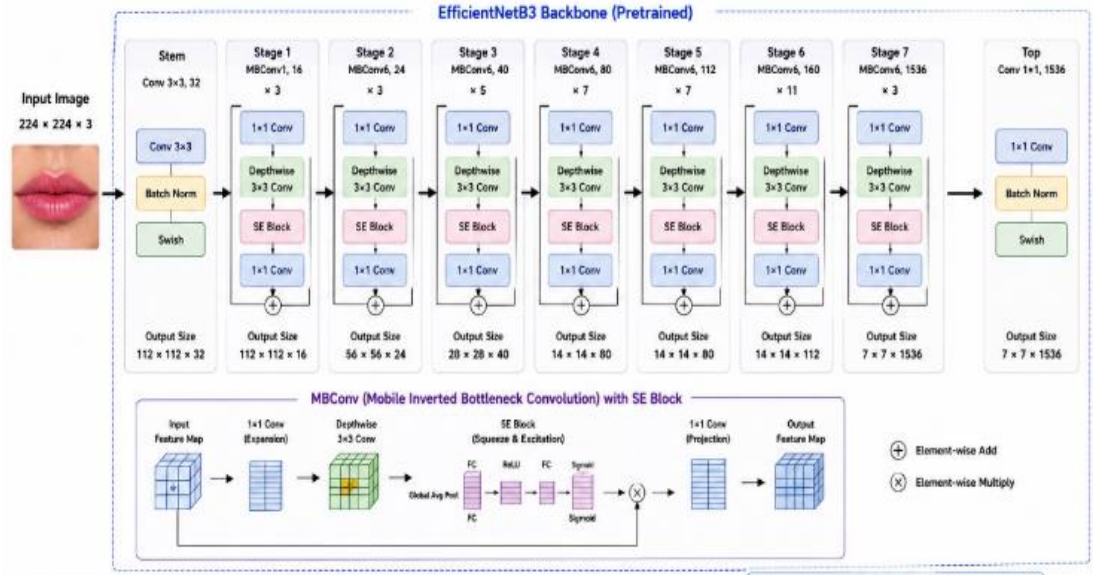

Fig.4: Efficientnetb3-Based Feature Extraction Architecture network

Fig.5 shows the classification layers of the Enhanced CNN. The attention-refined feature map obtained from the previous stage is first subjected to Global Average Pooling, producing a compact $1 \times 1 \times 1536$feature vector. This representation is passed through a sequence of fully connected layers with Batch Normalization, ReLU activation, and Dropout to learn discriminative features and reduce overfitting. The final fully connected layer maps the learned representation to the required vitamin classes, and a Sigmoid activation generates the corresponding vitamin deficiency probability scores for Vitamin A, Vitamin B, Vitamin C, and Vitamin D.

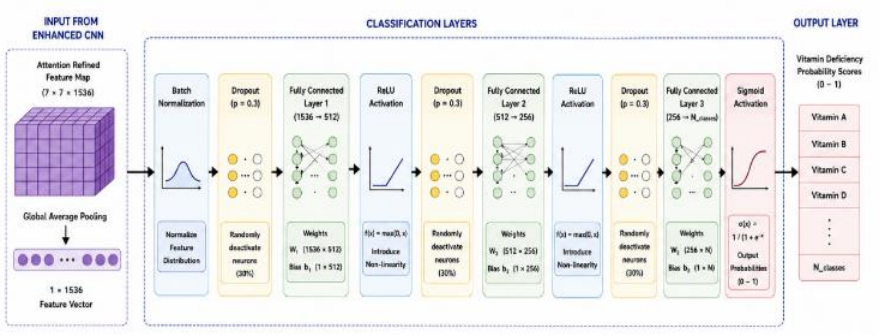

Fig.5: Classification Layer of Enhanced CNN network

Vision Transformer

The pre-processed image is divided into fixed-size patches and transformed into patch embeddings. For an input image of 224×224pixels with 16×16patches, the image produces $14 \times 14 = 196$patches. The self-attention operation is expressed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (5)$$

where Q, K, and Vrepresent query, key, and value matrices, respectively, and $d_k$is the key dimension. The resulting representation is used to generate the ViT prediction score S_{ViT}.

Swin Transformer

The Swin Transformer extracts hierarchical features using window-based multi-head self-attention (W-MSA) followed by shifted-window multi-head self-attention (SW-MSA). The attention operation within a window can be represented as

$$\text{W} - \text{MSA}(X) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \qquad (6)$$

where Brepresents the relative position bias. In the subsequent layer, the windows are shifted to establish connections between neighboring windows:

$$X^{l+1} = \text{SW} - \text{MSA}\big(\text{LN}(X^l)\big) + X^l \qquad (7)$$

The extracted hierarchical representation is used to obtain the Swin Transformer score S_{Swin}. Fig.6 shows the Swin Transformer-Tiny (Swin-T) architecture used in the proposed methodology. The $224 \times 224 \times 3$input image is first divided into 4×4patches and converted into patch embeddings through linear projection and position embedding. The embeddings are then processed through four hierarchical stages. Each stage uses Layer Normalization, Window-based Multi-Head Self-Attention (W-MSA), Shifted Window-based Multi-Head Self-Attention (SW-MSA), MLP layers, and patch merging to progressively reduce spatial resolution while increasing feature dimensions. W-MSA captures local relationships within windows, whereas SW-MSA shifts the windows to establish connections between neighboring regions. Finally, Global Average Pooling and a fully connected classification layer with Softmax generate the probability scores for the vitamin classes. This hierarchical structure enables the model to capture both local and broader visual patterns from the input image

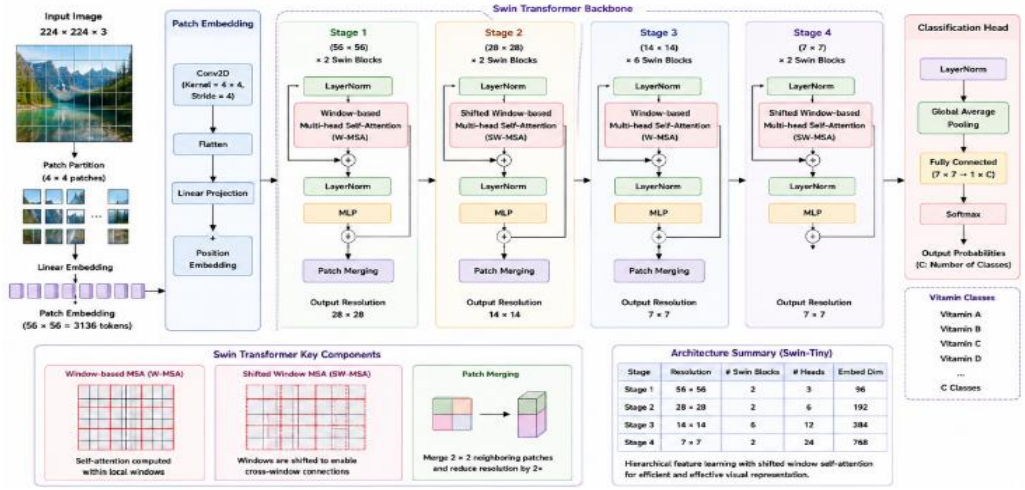

Fig.6: Swin Transformer-Tiny (Swin-T) architecture

Weighted Ensemble Fusion

The prediction scores obtained from the three models are combined using weighted ensemble fusion. According to the proposed architecture, the weights are 35% for CNN, 25% for ViT, and 40% for Swin Transformer. Therefore, the ensemble score is calculated as

$$S_E = 0.35S_{CNN} + 0.25S_{ViT} + 0.40S_{Swin} \qquad (8)$$

where $S_E$represents the final ensemble score.

Threshold-Based Decision

The ensemble score is compared with a predefined threshold T_b, where the threshold can be selected according to the corresponding body region. The classification rule is

$$C = \begin{cases} \text{Deficiency}, & S_E \geq T_b \\ \text{Healthy}, & S_E < T_b \end{cases} \quad (10)$$

where C represents the final classification and T_b represents the body-part-specific threshold.

Vitamin and Severity Identification

When the ensemble score satisfies the deficiency condition, the system identifies the corresponding vitamin category and severity based on the detected body region and prediction output. The final module provides the vitamin type, severity level, symptoms, precautions, and relevant food information. Thus, the proposed methodology combines local CNN features, global ViT features, and hierarchical Swin Transformer features to perform image-based preliminary vitamin deficiency screening.

3.2 Algorithm

Adaptive Multi-Scale Ensemble Algorithm for Vitamin Deficiency Screening

Input: Pre-processed image I

Output: Health status C, vitamin category V, severity S_v

1. Image Standardization:
   Resize the input image to 224×224, enhance contrast and illumination using CLAHE, and normalize the resulting image.
2. Parallel Feature Learning:
   Process the standardized image simultaneously through three complementary feature-learning branches:
   - Enhanced EfficientNet-B3 with spatial attention for local features.
   - Vision Transformer (ViT) for global patch-level relationships.
   - Swin Transformer for hierarchical and shifted-window features.
3. Attention-Guided CNN Representation:
   Generate the spatial attention map

$$A = \sigma\left(\text{Conv}_{1\times1}(F)\right)$$

   and obtain the refined representation

$$F_{CNN} = F \odot A.$$

4. Global Representation Learning:
   Divide the image into patches and obtain the ViT representation using self-attention:

$$F_{ViT} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

5. Hierarchical Representation Learning:
   Extract hierarchical features from the Swin Transformer using window-based and shifted-window attention to obtain F_{Swin}.
6. Independent Prediction:
   Generate the individual deficiency scores:

$$S_{CNN} = f(F_{CNN}),$$
$$S_{ViT} = f(F_{ViT}),$$
$$S_{Swin} = f(F_{Swin}).$$

7. Confidence-Aware Score Fusion:
   Combine the complementary predictions using weighted fusion:

$$S_E = w_C S_{CNN} + w_V S_{ViT} + w_S S_{Swin}$$

   where

$$w_C + w_V + w_S = 1.$$

   For the current architecture:

$$w_C = 0.35,\ w_V = 0.25,\ w_S = 0.40.$$

8. Body-Region-Specific Decision:
   Select the corresponding decision threshold T_b according to the detected body region b.

$$C = \begin{cases} \text{Deficiency}, & S_E \geq T_b \\ \text{Healthy}, & S_E < T_b \end{cases}$$

9. Vitamin Mapping:
   If deficiency is detected, associate the body region with the corresponding vitamin category:

$$V = M(b)$$

   where M represents the body-region-to-vitamin mapping.
10. Severity Estimation:
    Determine the severity level from the final deficiency score and classification output.
11. Information Generation:
    Generate the corresponding vitamin information, symptoms, precautions, and food recommendations.
12. Final Decision:
    Return

_

3.3: Implementation

Fig.7 illustrates the overall workflow of the proposed vitamin deficiency detection system. The input image is first validated and preprocessed before being passed simultaneously to three feature-extraction models: Enhanced CNN, Vision Transformer, and Swin Transformer. The Enhanced CNN uses spatial attention to generate the CNN score, while ViT and Swin Transformer generate the ViT and Swin-T scores, respectively. These scores are combined through ensemble score fusion and evaluated using a predefined threshold. Based on the resulting score, the system classifies the image as healthy or deficiency-related. If a

deficiency is detected, the corresponding vitamin category and severity are identified, followed by generation of relevant information and the final prediction output.

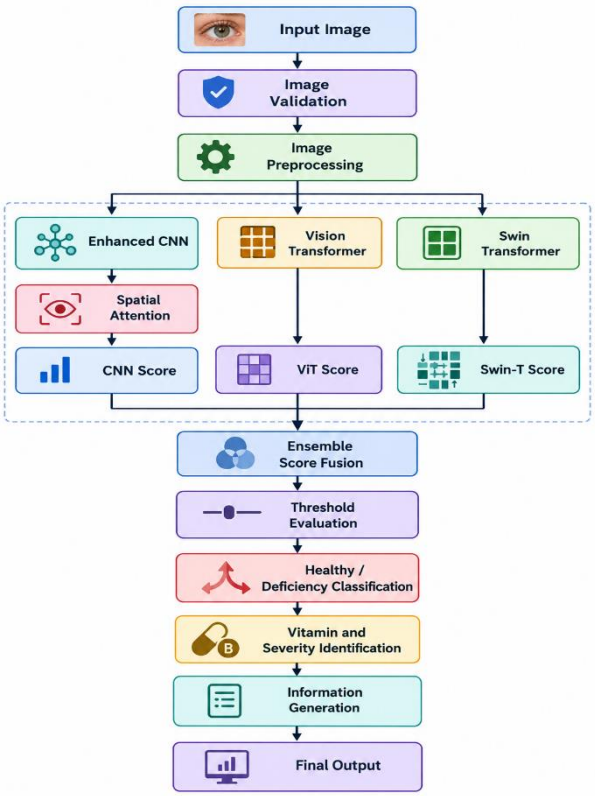

Fig.7: Implementation flowchart of the proposed method

.

IV. RESULTS AND DISCUSSIONS

Vitamin Deficiency Detection System was evaluated using images from five body regions: lips, eyes, nails, tongue, and skin. Each image was processed through the Enhanced CNN, ViT, and Swin Transformer branches, followed by ensemble score fusion and threshold-based classification. The system generated both deficiency-related and healthy-condition outputs, together with model scores, thresholds, vitamin categories, and severity levels.

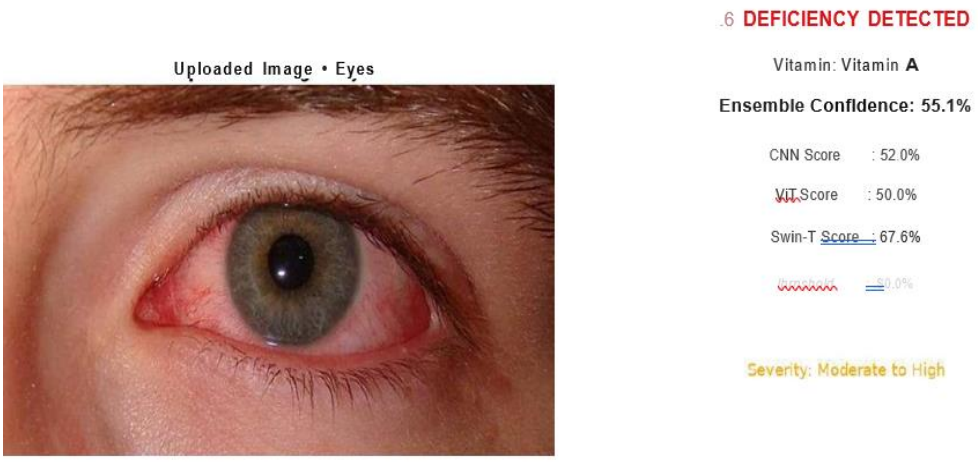

Fig.8: Vitamin A Deficiency Detection from Eye Image

The fig.8 shows the prediction for an eye image. The proposed system identifies Vitamin A deficiency with an ensemble confidence of 55.1%, exceeding the 50.0% threshold. The CNN, ViT, and Swin-T scores are 52.0%, 50.0%, and 67.6%, respectively, and the reported severity is Moderate to High. The fig.9 presents the result for a healthy lip image. The system produces a No Deficiency classification, with CNN and ViT scores of 0.0% and a Swin-T score of 34.2%, which is below the 47.0% threshold. The displayed healthy-condition confidence is 95.0%

Fig.9: Healthy Lip Image Classification

The fig.10 illustrates the classification of a nail image as Vitamin D deficiency. The CNN, ViT, and Swin-T scores are 47.4%, 50.0%, and 74.8%, respectively. The ensemble confidence of 55.3% exceeds the 42.0% threshold, resulting in a High severity classification

Fig.10: Vitamin D Deficiency Detection from Nail Image

The fig.11 shows the prediction obtained from a tongue image. The system detects Vitamin B deficiency with an ensemble confidence of 57.5%, compared with a 40.0% threshold. The CNN, ViT, and Swin-T scores are 45.7%, 50.0%, and 86.1%, respectively, and the reported severity is Moderate to High.

Fig.11: Vitamin B Deficiency Detection from Tongue Image

The fig.12 presents the classification of a healthy tongue image. The CNN and ViT scores are 0.0%, while the Swin-T score is 24.1%, which is below the 40.0% threshold. Therefore, the system produces a No Deficiency result with a displayed healthy-condition confidence of 95.0%.

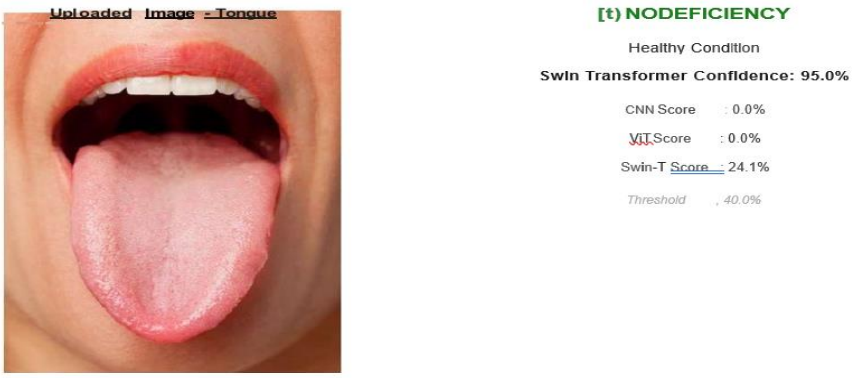

Fig.12: Healthy Tongue Image Classification

The fig.13 shows the result for a healthy eye image. The CNN and ViT scores are 0.0%, and the Swin-T score is 30.7%, remaining below the 50.0% threshold. The system consequently classifies the image as No Deficiency and reports a healthy-condition confidence of 95.0%.

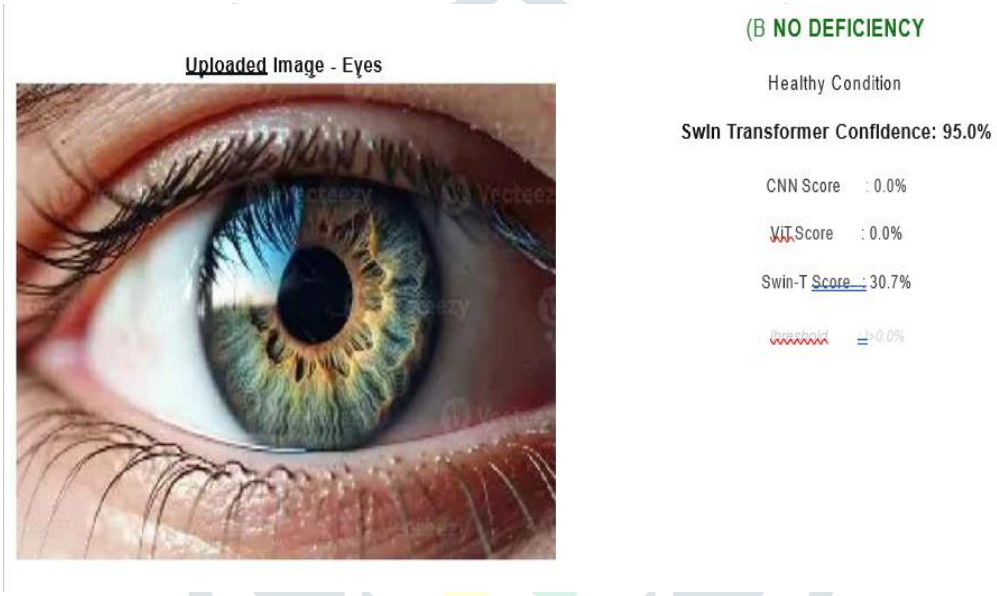

Fig.13: Healthy Eye Image Classification

The fig.14 presents the result for a healthy nail image. The CNN and ViT scores are 0.0%, while the Swin-T score is 23.6%, which is below the 42.0% threshold. The system therefore generates a No Deficiency classification with a displayed healthy-condition confidence of 95.0%.

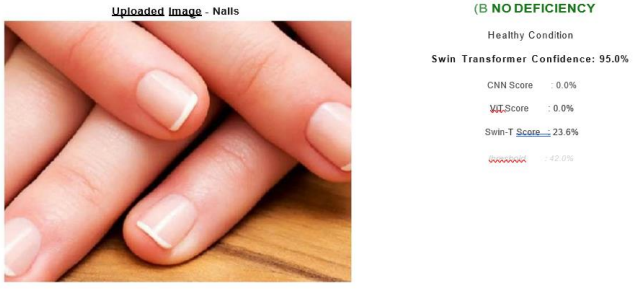

Fig.14:Healthy Nail Image Classification

The fig.15 illustrates the classification of a skin image as Vitamin A deficiency. The CNN, ViT, and Swin-T scores are 50.8%, 50.0%, and 67.3%, respectively. The ensemble confidence reaches 54.6%, exceeding the 41.0% threshold, and the system reports Moderate to High severity

.

Fig.15:Vitamin A Deficiency Detection from Skin Image

The fig.16 presents the result for a healthy skin image. The CNN and ViT scores are 0.0%, while the Swin-T score is 37.1%, which remains below the 47.0% threshold. The system consequently produces a No Deficiency classification with a displayed healthy-condition confidence of 95.0%

Fig.16: Healthy Skin Image Classification

Table 3: Overall Simulation Results of the Proposed Vitamin Deficiency Detection System

Input Region	Classification	Vitamin	Ensemble Confidence	CNN	ViT	Swin-T	Threshold	Severity
Lips	Deficiency	Vitamin C	50.8%	49.1%	37.3%	74.7%	47.0%	Moderate
Eyes	Deficiency	Vitamin A	55.1%	52.0%	50.0%	67.6%	50.0%	Moderate to High
Lips	No Deficiency	—	95.0%	0.0%	0.0%	34.2%	47.0%	Healthy
Nails	Deficiency	Vitamin D	55.3%	47.4%	50.0%	74.8%	42.0%	High
Tongue	Deficiency	Vitamin B	57.5%	45.7%	50.0%	86.1%	40.0%	Moderate to High
Tongue	No Deficiency	—	95.0%	0.0%	0.0%	24.1%	40.0%	Healthy
Eyes	No Deficiency	—	95.0%	0.0%	0.0%	30.7%	50.0%	Healthy
Nails	No Deficiency	—	95.0%	0.0%	0.0%	23.6%	42.0%	Healthy
Skin	Deficiency	Vitamin A	54.6%	50.8%	50.0%	67.3%	41.0%	Moderate to High
Skin	No Deficiency	—	95.0%	0.0%	0.0%	37.1%	47.0%	Healthy

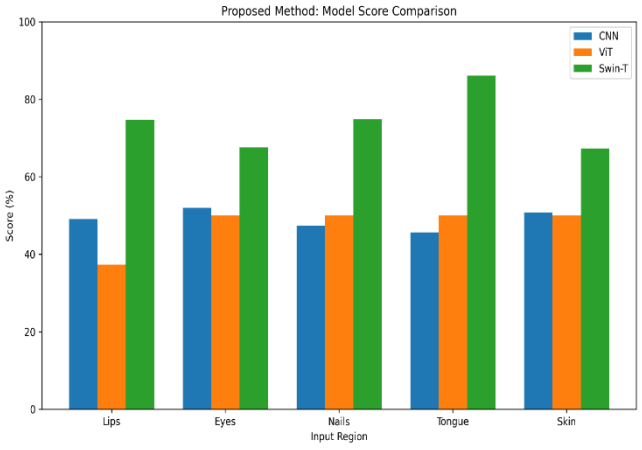

Fig.17: Proposed Method: Model Score Comparison

Fig.17 compares the CNN, ViT, and Swin-T scores for five input regions: lips, eyes, nails, tongue, and skin. The Swin Transformer produces the highest score for all evaluated regions, with the maximum value of 86.1% for the tongue. The CNN scores range from 45.7% to 52.0%, while the ViT scores range from 37.3% to 50.0%. The comparison illustrates the different responses of the three models to the evaluated input regions. Fig.18 compares the ensemble confidence with the corresponding threshold for the five deficiency-related input regions. The ensemble confidence is higher than the threshold for all five cases: lips (50.8% vs. 47.0%), eyes (55.1% vs. 50.0%), nails (55.3% vs. 42.0%), tongue (57.5% vs. 40.0%), and skin (54.6% vs. 41.0%). This comparison illustrates the threshold-based decision mechanism used for the reported deficiency classifications.

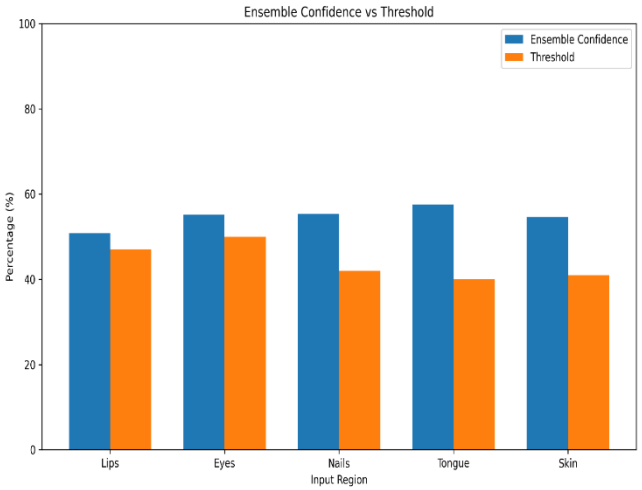

Fig.18: Proposed Method: Ensemble Confidence versus Threshold

V. CONCLUSION AND FUTURESCOPE

This paper presented a hybrid deep-learning framework for image-based preliminary vitamin deficiency screening using an Enhanced CNN, Vision Transformer (ViT), and Swin Transformer. The Enhanced CNN extracts local spatial and texture features, while ViT captures global relationships and Swin Transformer learns hierarchical visual representations through shifted-window attention. Their prediction scores are combined using an ensemble score-fusion mechanism followed by threshold-based classification. Simulation results for skin, eyes, lips, nails, and tongue demonstrated the operation of the proposed framework for both deficiency-related and healthy samples. Among the evaluated deficiency samples, the Swin Transformer achieved an average score of 74.1%, compared with 49.0% for CNN and 47.5% for ViT, with the highest Swin-T score of 86.1% obtained for the tongue sample. The system also generated vitamin-category and severity information for the tested deficiency cases. These results demonstrate the potential of combining complementary CNN and Transformer representations for preliminary image-based vitamin deficiency screening. Future work can focus on developing a larger and more diverse image dataset covering different age groups, skin tones, lighting conditions, and deficiency categories. The framework can be extended to include additional vitamins and multiple deficiency conditions. Adaptive or confidence-based ensemble fusion can be investigated instead of fixed model weights to dynamically exploit the most informative model for each input. Explainable AI techniques such as Grad-CAM and attention visualization can also be incorporated to identify the image regions influencing the prediction. Further work can explore real-time implementation on mobile, web, and edge-computing platforms, followed by evaluation on independent datasets and clinical validation to determine the practical reliability of the system

REFERENCES

[1] G. Sambasivam, J. Amudhavel, G. Sathya, “A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods”, published in IEEE Access, vol 8, pp. 109492 - 109507.

[2] E. K and S. K, "Diagnosis of Vitamin Deficiency in Human Beings using DNN Algorithm," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1627-1632'.

[3] Gul, Zeki, and Sebnem Bora. “Exploiting Pre-Trained Convolutional Neural Networks for the Detection of Nutrient Deficiencies in Hydroponic Basil.” Sensors (Basel, Switzerland) vol. 23,12 5407. 7 Jun. 2023.

[4] 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.

[5] S.Solanki, U.P.Singh, S.S.Chouhanand, S.Jain, "Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview," in IEEE Access, vol. 11, pp. 12870-12886, 2023

[6] Rahman, Takowa, and Md Saiful Islam." MRI brain tumor detection and classification using parallel deep convolutional neural networks." Measurement: Sensors 26 (2023): 100694.

[7] Patil, Suraj, and Dnyaneshwar Kirange. "Ensemble of deep learning models for brain tumor detection." Procedia Computer Science 218 (2023): 2468-2479.

[8] Satyanarayana,Gandi,P.AppalaNaidu,VenkataSubbaiah Desanamukula, and B. Chinna Rao. "A mass correlation based deep learning approach using deep convolutional neural network to classify the brain tumor." Biomedical Signal Processing and Control 81 (2023): 104395.

[9] Gupta, BrijB., Akshat Gaurav, and Varsha Arya."Deep CNN based brain tumor detection in intelligent systems." International Journal of Intelligent Networks 5 (2024): 30-37.

[10] Akter, Atika, et al. "Robust clinical applicable CNN and U-Net based algorithm for MRI classification and segmentation for brain tumor." Expert Systems with Applications 238 (2024): 122347.

[11] Kumar, Sunil, and Dilip Kumar. "Human brain tumor classification and segmentation using CNN." Multimedia Tools and Applications 82.5 (2023): 7599-7620. [

[12] Mohsen, Saeed, et al. "Brain Tumor Classification Using Hybrid Single Image Super-Resolution Technique with ResNext101_32x8d and VGG19 Pre-Trained Models." IEEE Access (2023).

[13] Rajendran, Surendran, etal. "Automated segmentation of brain tumor MRI images using deep learning." IEEE Access (2023).

[14] Jabbar, Ayesha, et al. "Brain tumor detection and multi-grade segmentation through hybrid caps-VGGNet model." IEEE Access (2023).

[15] Abhijit Ashok Hipparkar, "Vitamin Deficiency Detection Using Image Processing and Neural Network" Member, Computer Engineering Department, College of Engineering, Phaltan, Maharashtra, India. Jul 15, 2023.
[16] Swapnil Bonde, Pravin Argade, Prof. Deepali M. Gohil, Rutik Gagare, Ganesh Wakade, "Vitamin Deficiency Detection using Image Processing and Deep-CNN Algorithm", International Research Journal of Modernization in Engineering Technology and Science, vol. 6, Issue 05, May 2024.
[17] M Jayaram, T. Nikhitha, Vasa Mani Shankar, Kanduri Nandini, Dommata Swetcha, Sudhanaboina Sai Sriman, "Deep Learning Technique Dense-Net Application to Predict Vitamin Deficiency", Industrial Engineering Journal, vol. 53, issue 5, pp.118-129, May 2024.