AGRI MAGAZINE
(International E-Magazine for Agricultural Articles)
Volume: 03, Issue: 08 (August, 2026)
Available online at http://www.agrimagazine.in
©Agri Magazine, ISSN: 3048-8656

# Application of Vision Transformer (ViT) for Wheat Yield Prediction

**B.K. Singh[1], *Dulal Chakrabarti[2], Jaya Singh[1] and Akul Kumar[1]**
[1]BKC WeatherSys Pvt. Ltd., H-135, Sector 63, Noida 201307, India
[2]India Meteorological Department (Retired), New Delhi - 110003, India
*Corresponding Author's email: dulalchakrabarti@gmail.com

Machine Learning (ML) models have emerged as a transformational technology in agriculture. International Food Policy Reasearch Institute (IFPRI) commissioned BKC WeatherSys Pvt. Ltd. to study the use of smart phone pictures for wheat yield prediction in 2017. From the detailed report [Singh et al (2018)] it may be seen that Convolution Neural Network (CNN) Model provided encouraging results of accuracy above 50%. Vision Transformer (ViT) model was introduced in 2021 [Dosovitskiy et al. in (2021)] and a pre-trained version named ImageNet-21k provided orders of magnitude improvement. India has a widespread mobile network infrastructure (PIB 2025). In this paper we explore the application of a pre-trained Vision Transformer (ViT) to the same IFPRI image dataset to achieve accuracy of over 80%.

## Introduction

A detailed review of ViT is available in the literature [Dosovitskiy et al. (2021)]. The core equation for the scaled dot-product attention in a single head is expressed as (Equation 1):

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\text{sqrt}(d_k)) \text{ dot } V \quad \text{.....Equation 1.}$$

The crop field image is first split into non-overlapping patches (e.g., 16 x 16 pixels), which are then linearly embedded into vectors (tokens). These tokens, along with a special (CLS - Classification) token for classification, form the input sequence. Query (Q) calculates the context-aware representation of the current patch by asking - "What information am I looking for?" The key (K) represents all other patches (including itself) answers "This is the information I contain" and the value (V) represents the actual content or feature information of all other patches that is aggregated. The 'attention weights" are what the heat maps show and determine how much focus the querying patch (Q) places on every other patch (K) in the image. The dot product between Q (the current patch) and all K's (all patches) computes the similarity or relevance score between them. A high score means the two patches contain relevant information to one another. The scores are divided by the square root of the dimension of the key vectors. This stabilizes the gradients during training, preventing the dot products from becoming too large when the dimension $d_k$ is high. The scaled scores are normalized using the SoftMax function. This turns the scores into a set of attention weights (probabilities) that sum to 1. The resulting heatmap visually represents these weights: Bright/hot areas mean high weight or strong attention and Dark/cool areas mean low weight or weak attention. The attention weights are multiplied by the corresponding Value vectors (V) to create a new, contextualized vector for the querying patch. This vector now incorporates information from the entire image, weighted by relevance, which is the core strength of the Transformer. The multi-head part means, this process is run independently in parallel across a number of "heads" (12 in the base model). Each head learns different Q, K, V linear projections, allowing it to focus on different types of patterns. The outputs of all

heads are then concatenated and linearly combined to learn visual patterns by making the attention weights meaningful for the final classification task (wheat yield).

## Global Context and Long-Range Dependencies

Unlike Convolutional Neural Networks (CNNs), which use local receptive fields to build up context hierarchically, the self-attention mechanism in ViT inherently models global context from the first layer. By attending to every other patch, the model learns long-range dependencies which is crucial for tasks like yield prediction, where field-wide consistency is a key pattern. In our heat maps, a single patch might be querying (itself) attending strongly to a specific textural area while ignoring a sparse or barren area. This highlights the learned pattern of a relevant feature. Each attention head develops a specialized function over the multiple Transformer layers (Table 1):

**Table 1: Crop image multiheaded attention layers**

| Attention Head Function | Learned Pattern | Relevance to Crop Image |
|---|---|---|
| Early Layers (Low-Level) | Geometric/Local features (edges, corners, lines) | Attending to boundaries between fields or plants. |
| Middle Layers (Mid-Level) | Texture and Object Parts (e.g., specific leaf shapes) | Attending to textural uniformity or patterns of disease/damage on a crop. |
| Late Layers (High-Level) | Semantic Information and Global Features | Attending to all patches that constitute the entire "wheat crop" object in the image, or features highly correlated with the final yield label. |

## The Role of Pre-training (ImageNet-21k)

This pre-training allows the ViT to learn a highly general set of visual features—edges, textures, object parts, and compositions—from a vast array of natural images. When we fine-tune this model for wheat yield classification, the attention heads are adjusted to leverage those general patterns and focus on the specific agricultural features that are most predictive of the target (e.g., density, color uniformity, presence of disease, or water stress signatures). The heat maps show that a specific patch's feature is being compared to, and aggregated with, the features of other highly relevant patches, forming a spatially weighted "mental image" that is passed on to the next layer for a more accurate classification.

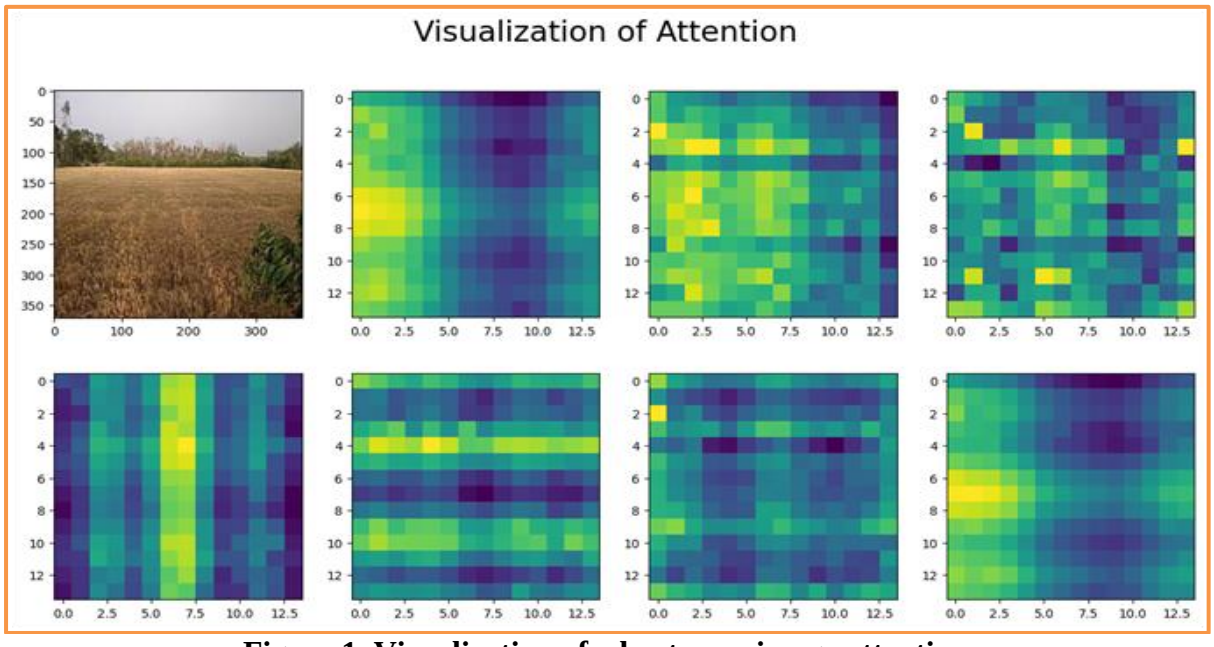


**Figure 1. Visualization of wheat crop image attention**

## Image Data and Methodology of Pre-processing and Processing

We examined a total of 5818 smart phone pictures reported by 637 farmers from November 2016 to April 2017. The farmers were identified by a unique 6-digit census village code and plot numbers. However, out of 637 farmers, only the yields from the plots of 437 farmers were verified by crop cutting; hence only the pictures reported by these 437 farmers were used for training and evaluation of the ViT model. The picture database was organized in the following hierarchy (Figure 2).

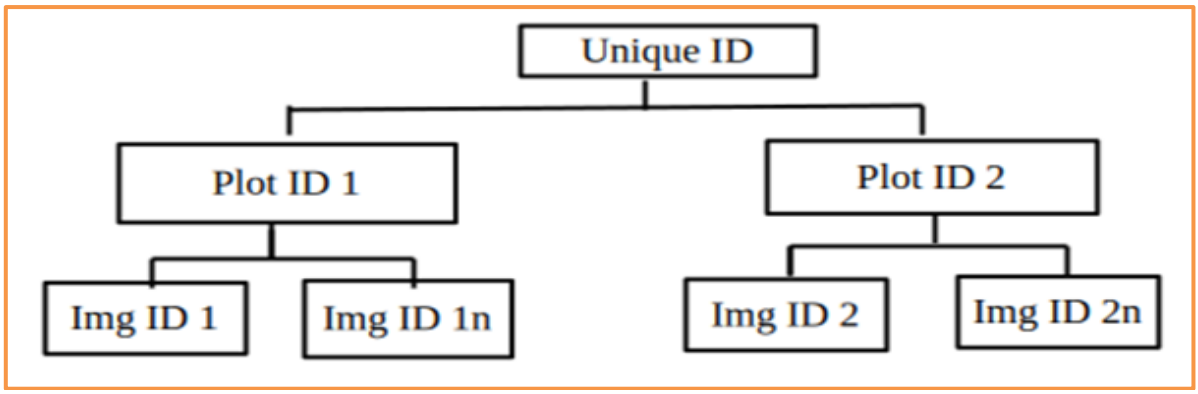


Figure 2. Hierarchy of picture database

The farmers in the field used a mobile app named "Wheatcam" to upload their pictures to the IFPRI server, which then stored the files in the hierarchy shown in Figure 2 and added an entry into an index file used for accessing data from the database. with a standard file naming convention (as shown below).

dddddd_n_sssssssssssss

where dddddd is the 6-digit unique id, n is the single-digit site id, and sssssssssssss is the 13-digit time stamp in milliseconds. Every picture passed 5 processing stages: 1) reading data, 2) pre-processing data, 3) train the pre-trained ViT model, 4) fine tune to avoid overfitting of model parameters, and 5) validate model prediction. These steps are illustrated in the following schematic diagram (Figure 3).

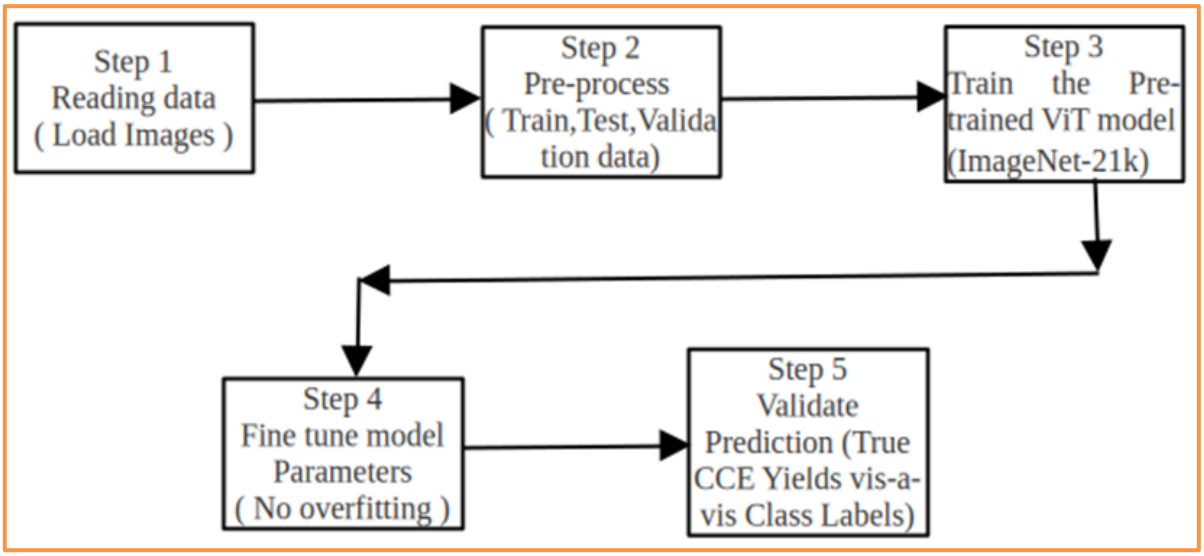


**Figure 3. ViT Process Schematic**

The raw input images are uploaded and extracted into the working directory. The images and their associated labels inferred from folder structure are loaded into a Hugging Face Dataset object. The full dataset is partitioned into two distinct sub-sets: Training (for model updates) and Validation for performance evaluation during training and for final unbiased evaluation. The ViT-specific pre-processing configuration (for google/vit-base-patch16-224-in21k) is loaded, including parameters for resizing, scaling, and normalization. The core pre-processing ensures all input images are in the required RGB format. It applies the pre-trained

ViT transformations including resizing the image to 224x224 pixels, normalizing the pixel values and converting the processed image into a PyTorch tensor format (pixel values).The datasets are now ready, containing processed tensors (pixel values) and corresponding class labels (Figure 4).

```
DatasetDict({
    train: Dataset({
        features: ['image', 'label'],
        num_rows: 4654
    })
    validation: Dataset({
        features: ['image', 'label'],
        num_rows: 1164
    })
})
Train size: 4654
Validation size: 1164
```

**Figure 4. Image Dataset**

The pre-trained weights from ViT's base model (trained on a massive dataset like ImageNet-21k) are loaded. The classification head is customized for the new task by setting 5 classes. Wheat yields were classified as - less than 10, 10 to 15 ,15 to 20, 20 to 25, more than 25 quintals per hectare (Table 2).

**Table 2. Wheat Yield Classes in Quintals and Labels**

| Less than 10 | 10-15 | 15-20 | 20-25 | More than 25 |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

Hyperparameters (e.g., per device train batch size=8, train epochs=2) initialize the Hugging Face Trainer object - linking the model arguments and the prepared Training and Validation datasets. The pre-trained ViT weights are fine-tuned to recognize the new, specific classes (wheat characteristics) by minimizing the loss function on the Training Set. The evaluation strategy helps avoid over-fitting by periodically checking performance on the Validation Set.

**Table 3. Training Progress and Performance**

| Metric | Value | Notes |
|---|---|---|
| Global Step | 1746 | Total batches processed. |
| Epoch | 3 | Two full passes through the training data have been completed. |
| Training Loss | 0.2923 | The low loss value indicates the model is fitting the training data very well. |

**Table 4. Resource Usage and Speed Metrics**

| Metric | Value | Unit |
|---|---|---|
| Training Runtime | 281.3681 | seconds |
| Samples per Second | 49.622 | samples/sec |
| Steps per Second | 6.205 | steps/sec |
| Total FLOPs | $1.081*10^{18}$ | floating-point operations |

The trained model generates class-specific logits (raw, unnormalized predictions) for every image in the validation set. The true labels and the final predicted class indices are extracted; overall model performance is measured using metrics like Precision, Recall, and F1-Score for each of the 5 image classes as shown under Table 5.

**Table 5. Model Performance**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 10 | 0.87 | 0.79 | 0.83 | 75 |
| 15 | 0.81 | 0.74 | 0.77 | 193 |
| 20 | 0.78 | 0.88 | 0.83 | 366 |

| | | | | |
|---|---|---|---|---|
| 25 | 0.83 | 0.85 | 0.84 | 342 |
| 30 | 0.84 | 0.71 | 0.77 | 188 |
| | | | | |
| Accuracy | | | 0.81 | 1164 |
| Macro Avg | 0.82 | 0.79 | 0.81 | 1164 |
| **Weighted Avg** | **0.81** | **0.81** | **0.81** | **1164** |

A Confusion Matrix (Figure 5) is generated and plotted, providing a visual breakdown of where the model is making correct versus incorrect predictions (e.g., predicting class "10" when the true class was "15").

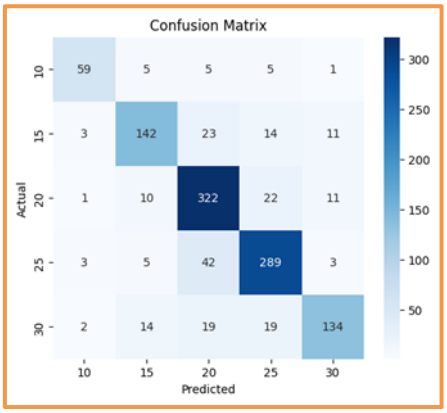


**Figure 5. Confusion Matrix**

The final output (Figure 6) is an image class prediction which can be used for deployment or further analysis by visually comparing the model's prediction with the true label on sampled images.


Figure 6. Final Prediction of yield class

## Results and Discussion

The model shows an accuracy of 81% and a balanced overall performance with an F1-Score of 0.81. While the performance is strong, there's room for improvement, particularly in identifying true positives for the minority classes (above 25 quintals per hectare). Based on the classification report, here is a detailed analysis and rating of the wheat image classification model. The total number of samples is 1164. Classes '20' and '25' have 366 and 342 samples (about 31% and 29% of the data). Classes '10' (75 samples) is under-represented. When a dataset is imbalanced, metrics like high Accuracy (here, 0.81) can be misleading, as the model might be predicting false alarm. So, a balanced dataset would be the ideal requirement.

## Concluding Remarks

Starting with the IFPRI project experiment at BKC WeatherSys Pvt. Ltd. in 2017 a full stack CNN model in the form of metGIS-Agro at the backend serving the smart phone frontend using "FasalSalah" mobile App has been in operation all along providing an operational framework to upgrade it with data driven Vision Transformer model.

## Acknowledgment

The authors acknowledge with thanks M/s BKC WeatherSys Pvt. Ltd (https://weathersysbkc.com/ ) and also IFPRI (International Food Policy Research Institute) (Singh, B. K., et al. (2018)) for providing with data and support to this work.

## Disclaimer

The contents and views expressed in this article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

## References

1. Singh, B. K., Chakraborty, Dulal, Kalra, Naveen, and Singh, Jaya (2018), A Tool for Climate Smart Crop Insurance: Combining Farmers' Pictures with Dynamic Crop Modelling for Accurate Yield Estimation Prior to Harvest https://cgspace.cgiar.org/items/d5be41df-750d-4325-8300-f77da3249f70
2. Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, Houlsby Neil (2021), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://arxiv.org/abs/2010.11929
3. PIB (Press Information Bureaue) survey report (2025), https://www.pib.gov.in/PressReleasePage.aspx?PRID=2132330
4. Berroukham Abdelhafid, Housni Khalid and Lahraichi Mohammed (2023), Vision Transformers: A Review of Architecture, Applications, and Future Directions, 7th IEEE Congress on Information Science and Technology (CiSt)
5. Yaoli Wang, Deng Yaojun, Zheng Yuanjin, Chattopadhyay Pratik and Wang Lipo (2025) Vision Transformers for Image Classification: A Comparative Survey, *Technologies 13*(1),2; https://doi.org/10.3390/technologies13010032