

Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 40
UAV and ML Applications in Agriculture (Contd.)

(Refer Slide Time: 00:22)

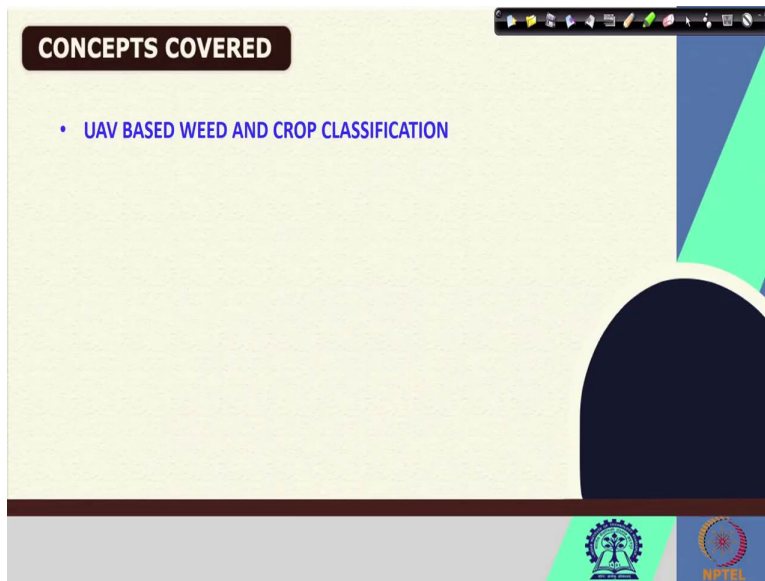


Welcome, friends, to this 5th lecture of week 8 of NPTEL online certification course of Machine Learning for Soil and Crop Management. And in this week, we are discussing different types of deep learning methods, and, like recurrent neural network, how it operates, and then we have discussed about Vision Transformer, then self-attention and how this Vision Transformer works, we have discussed in our lectures, previous lectures.

Also we have discussed what are the benefits of using UAVs in agricultural operations, what are the different types of applications in agriculture for UAVs, what are the different types of UAVs, what are the steps of UAV-based operations in agriculture. So, in this lecture, last lecture, we are going to discuss, we are going to focus on one paper which has been recently published, and I will share that paper, that is open access paper, So, I will share that paper in our course platform.

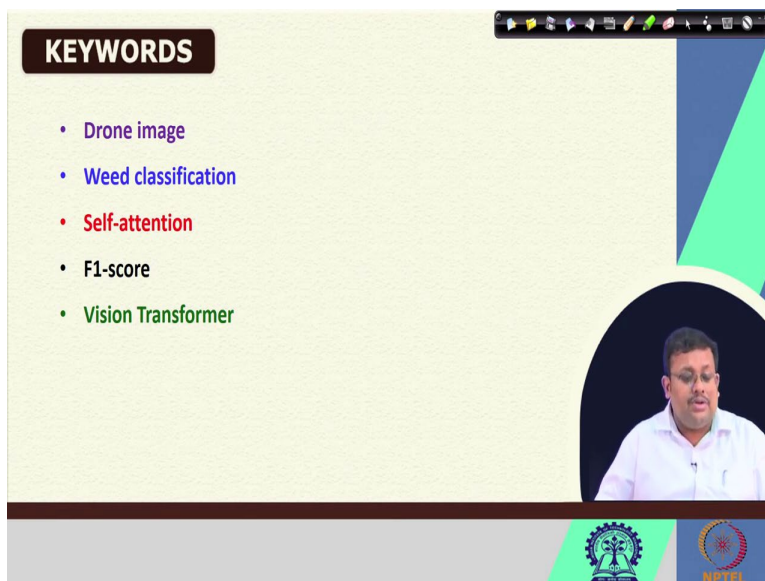
And in this paper they have used the UAV based images for, and also different types of machine learning algorithms, for crop and weed classification. We are going to see, by using different types of calibration validation strategies, we are going to also see that.

(Refer Slide Time: 01:55)



So, basically, the concept which we are going to discuss in this lecture is UAV based weed and crop classification. So, this is the major topic of this lecture.

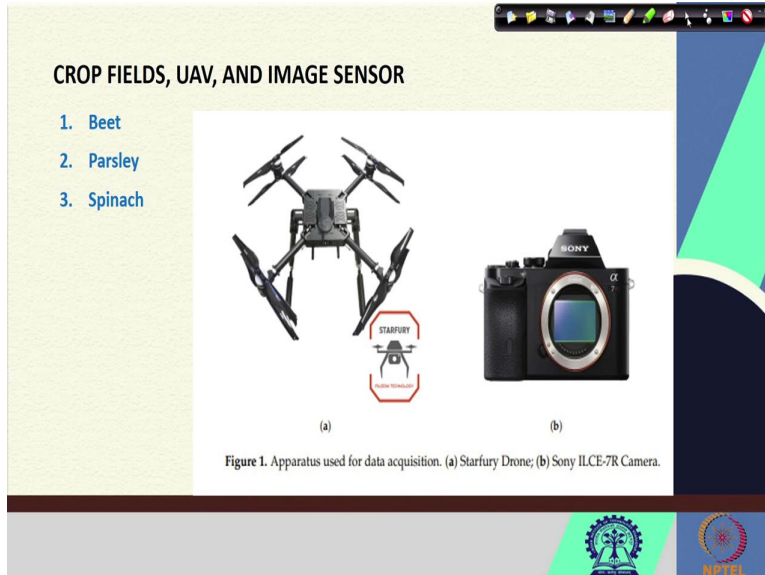
(Refer Slide Time: 02:04)



So, these are the keywords of this lecture. We are going to discuss about the drone images or UAV images, and then weed classification, then self-attention F1-score, and also Vision Transformer. We have already learned about the basics of Vision Transformer. So, we are going to see how they are applying the Vision Transformer in

real application where they are using the UAV based images of the crops and weed for their subsequent classification.

(Refer Slide Time: 02:33)



Now, in this paper, Reedha et al, in this paper, they have used they have used the images for three different types of crops. One is, they have focused first on beet, and also they have focused on parsley and spinach. And they have used, let me just increase the font, So, they have used the, this STARFURY drone, which is a quadcopter. And also they have used this Sony ILCE-7R camera, which combined with this STARFURY drone, and they have taken the images of the crop fields from certain elevation.

(Refer Slide Time: 03:31)

IMAGE ACQUISITION

1. 30 m = beet field
2. 20 m = parsley and spinach field
3. These altitudes were selected to minimize drone flight times while maintaining sufficient image quality






Figure 1. Apparatus used for data acquisition. (a) Starbury Drone; (b) Sony ILCE-7R Camera.



So, they have fixed the elevation of 30 meter for beet field. And for parsley and spinach, they have fixed the elevation at 20 meters. So, they have taken the image from 30 meter distance and 20 meter distance from beet field and parsley and spinach field. And these altitudes were selected to minimize the drone flight times while maintaining the sufficient image quality. So, depending on the crop, they have adjusted the elevation from where this UAV will take the image using this Sony camera, and then to ensure that minimum flight time is maintained while also ensuring the sufficient image quality.

(Refer Slide Time: 04:22)

IMAGE ACQUISITION

- The drone followed a specific flight plan and the camera captured RGB images at regular intervals.
- The images captured have respectively a minimum longitudinal and lateral overlapping of 70% and 50–60% depending on the fields vegetation coverage and homogeneity, assuring a better and complete coverage of the whole field of 4 ha and improving the accuracy of the orthorectified image of the field.



Figure 2. Overlay of the orthophoto on google earth of the spinach plot (left) and the flight plan (right) across a spinach field (the images are taken along the yellow lines at regular intervals to ensure sufficient overlapping).

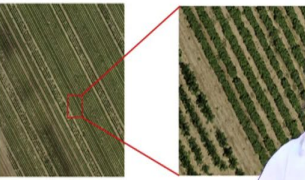




Figure 3. Example of image captured from a spinach study site.



So, using those images they have, using those configurations or set up, they have taken images. So, the drone followed a specific flight plan. As you can see in this picture, this is the flight plan they have used, and the camera captured these RGB images. So, simple RGB images they have captured at regular intervals. Now, remember that the images captured have respectively a minimum longitudinal and lateral overlapping of 70 percent and 50 to 60 percent depending on the vegetation coverage and also the homogeneity.

So, depending on the vegetation coverage of the field for these three crops as well as weeds, they have maintain a longitudinal and lateral overlapping of 70 percent and 50 to 60 percent to assure that a better and complete coverage of the whole field of this 4 hectares. So, the field is 4 hectare land. So, this 4 hectare land can be completely covered and improve the accuracy of the orthorectified image. So, what is orthorectification? We are going to learn.

So, you can see here, this first figure shows the overlaying the orthophoto on Google Earth of the spinach plots. So, here they have overlaid this orthophoto on the Google Earth. You all know about the Google Earth. So, this orthophoto has been overlaid on this Google Earth of the spinach plot, and then this is the flight plan as I have mentioned, across this spinach field, and the images are taken along these yellow lines. So, these are the flight plans. So, the drone move in this fashion, and the camera took the images at regular interval to ensure the sufficient overlapping.

So, this flight plan is ensuring that sufficient coverage or overlapping should be there So, that the image can be used to cover the whole area. So, while they have taken, it is an, you can see here, these are the examples of image captured from a spinach study site.

(Refer Slide Time: 06:50)

ORTHORECTIFIED IMAGE

- Raw aerial or satellite imagery cannot be used in a GIS until it has been processed such that all pixels are in an accurate (x,y) position on the ground.
- Photogrammetry is a discipline, for processing imagery to generate accurately georeferenced images, referred to as **orthorectified images** (or sometimes simply **orthoimages**).
- Orthorectified images have been processed to apply corrections for optical distortions from the sensor system, and apparent changes in the position of ground objects caused by the perspective of the sensor view angle and ground terrain.

<https://www.esri.com/about/newsroom/insider/what-is-orthorectified-imagery/>

So, once they have taken the images, they are talking about orthophoto, orthorectification. So, then, so, the question comes what is orthorectification, and what is orthorectified image? So, you can see that this image is a raw aerial or satellite image which can be taken by any sensor, either satellite sensor or aerial sensor at different orientation. So, remember, that raw aerial or satellite imagery cannot be used in a GIS until it has been processed such that all pixels are in accurate x and y position on the ground.

So, I have already told you about the photogrammetry. So, this photogrammetry is a discipline for processing the imagery to generate accurately geo-referenced images, and these reference geo-referenced images are known as orthorectified images or sometime we call it orthoimages.

In the last slide, I have showed you one orthoimage. So, this orthoimage, which has been taken by this UAV based Sony RGB camera, is orthorectified because when that drone is hovering over the field and taking the images of the crop, due to its different tilting angle the image has to be corrected. So, that they can be positioned in an accurate x-y position, they can be fixed in an accurate x-y position on the ground. So, remember, that this is called the orthorectification.

Now, orthorectified image have been processed to apply corrections for optical distortion from the sensor system and apparent changes in the position of the ground object caused

by the perspective of the sensor view angle and ground terrain. So, here, this is the raw photograph. You can see here the raw photograph is taken by, suppose there is a sensor here, and it is making a 25 degree angle. So, of course, this will be the raw image. And after we take this raw image, this tilting of the sensor will have an effect on this image.

And this orthorectification will take, will basically rectify the image by considering the view from this Nadir point, that is, from the top, from the top view of the image. So, in this way an image which is taken by any sensor, either it is a satellite or it is a aerial platform, these images can be rectified using this orthorectification. And this is very much necessary.

So, these orthorectified images were utilized in this study also. It is just for an example, in this slide, but remember that this orthorectification is important So, that we can get the best result out of our images.

(Refer Slide Time: 10:13)

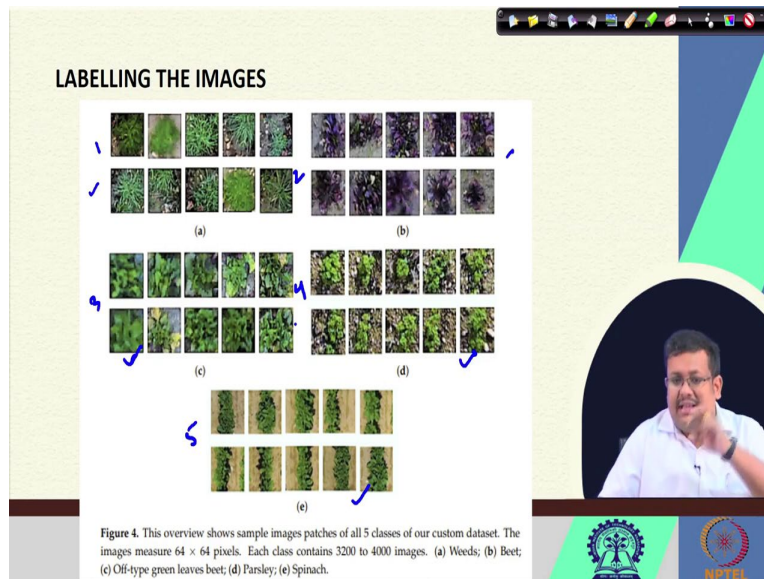
ORTHORECTIFIED IMAGE

- The orthorectification process requires: An accurate description of the sensor, typically called the sensor model; detailed information about the sensor location and orientation for every image; and an accurate terrain model

<https://www.esri.com/about/newsroom/insider/what-is-orthorectified-imagery/>

So, remember that, this orthorectification process requires certain things, an accurate description of the sensor, typically called the sensor model, and detailed information about the sensor location and orientation for every image, and an accurate terrain model. So, using these components, you can orthorectify an image.

(Refer Slide Time: 10:38)



Now, in this study, once these images were taken, now the next step was drawing some region of interest or boxes and from those, they are going, they have labeled the images. So, here, you can see that after taking the images it is very input, the next important step is, you should label the images of, and give it the proper classification.

So, here, you can see there are five classes, 1, 2, 3, 4 and 5. So, the first, the first class represent the images of the weeds, different types of weeds. So, the second class represents the images of beet plants, and the third class represented the images of the off type green leaves beet, and this is, the fourth one is parsley plants, and the fifth one is the spinach plant.

So, these classes, so, this basically, this overview shows the sample image patches of all the five classes of that custom dataset which they have created after taking the image using the UAV. So, once you take the image using the UAV, the next step is you should label the image. And remember that the images has a resolution of 64 by 64 pixels. So, all the images and the resolution of 64 by 64 pixels.

(Refer Slide Time: 12:25)

IMAGE PREPROCESSING

- Images have been rescaled to 0-1 range and then normalized by scaling the pixels values to have a zero mean and unit variance before being divided into training, validation and testing sets
- Data augmentation to improve the model robustness

Table 1. Class Distribution.

Class	Number
Weed	4000
Beet	4000
Off-type Beet	3265
Parsley	4000
Spinach	4000

Now, once the images are labeled, the next step is image processing. So, here, in the image processing steps what happens, images have been rescaled. So, those images which were taken for weeds, for beets, for off green type beet, and also the parsley and the spinach, they were processed to rescale for 0 to 1 range, and then normalized by scaling the pixel values to have a 0 mean and a unit variance before being divided into training, validation and testing sets.

So, the whole images were divided into training and testing and validation steps. We are going to discuss different validation strategies. And also they did some data augmentation to improve the model robustness. So, some images among this class, for example these, off type beets had very less number of images. So, they employed some strategy to augment the number of images.

And that is called data augmentation. And generally we go for the data augmentation specifically for image based analysis So, that we can maintain the balance between the classes to get the maximum accuracy from our classification algorithm. So, here also you can see we have got for 4,000 images for weed, 4,000 images for beet, 3,265 images for off type beet, for parsley we have also got 4,000, and for spinach we have got also 4,000 images.

(Refer Slide Time: 14:20)

RECALL: VISION TRANSFORMER (ViT)

- The Vision Transformer, or ViT, is a model for image classification that employs a Transformer-like architecture over patches of the image
- An image is split into fixed-size patches, each of them are then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder
- In order to perform classification, the standard approach of adding an extra learnable "classification token" to the sequence is used
- The self-attention enhances some parts of the input data while diminishing other parts — the thought being that the network should devote more focus to that small but important part of the data

The diagram, titled "Vision Transformers", illustrates the process. It starts with an input image (a landscape) being divided into patches. These patches are processed by a "Linear Projection of Flattened Patches" block, which outputs a sequence of vectors. These vectors are then fed into a "Transformer Encoder" block. The output of the encoder is passed to an "MLP Head" block, which produces the final classification result. A "CLS" token is shown as an input to the MLP Head. The diagram also includes a small inset image of a person speaking.

Credit: David Cozzomini (CC BY-SA 4.0)

So, once we got those images by augmentation and we have, they have rescaled those images and normalized those images, the next they have used the Vision Transformer as well as self-attention. So, from your previous lecture that in the vision transformer, a model which is a model for image classification that employs a transformer like architecture over the patches of the image.

So, here, the images are divided into several patches and those, and each of them were then linearly embedded, position embeddings were added, and then resulting sequence of the vectors are fed to the, are fed into the transformer encoder. And in order to perform this classification, the standard approach of adding an extra learnable classification token on the sequence is used.

And finally, through the process of self-attention, it enhances, it helps in proper classification. Now, what is the self-attention? If you recall, the self-attention basically a process which enhances some parts of the input data while diminishing the other part and basically the idea behind this self-attention is to, is that the network should devote more focus to the small but important part of the data, and less focus to the unimportant part of the data.

(Refer Slide Time: 15:45)

SELF-ATTENTION

- If each pixel in a feature map is regarded as a random variable and the covariances are calculated, the value of each predicted pixel can be enhanced or weakened based on its similarity to other pixels in the image
- The mechanism of employing similar pixels in training and prediction and ignoring dissimilar pixels is called the self-attention mechanism
- It helps to relate different positions of a single sequence of image patches in order to gain a more vivid representation of the whole image

Vision Transformers

Credit: David Coccomini (CC BY-SA 4.0)

The diagram illustrates the Vision Transformer architecture. It starts with an input image (a landscape with trees and a field) which is processed by a 'Linear Projection of Flattened Patches' block. The output is a sequence of patches, each represented by a small colored square. These patches are then fed into a 'Transformer Encoder' block. The output of the encoder is a sequence of tokens, each represented by a small colored square. Finally, these tokens are processed by an 'MLP Head' block to produce the final output.

So, we have discussed this self-attention in details in our previous lecture. So, I am not going to detail it further, but just remember that it helps to relate the different positions of a single sequence of image patches in order to gain a more vivid representation of the whole image. So, this is how this self-attention generally occurs in case of Vision Transformer.

(Refer Slide Time: 16:12)

SELF-ATTENTION

Original image **Attention map**

Figure 5. Attention mechanism on an image patch (left) containing weeds (in green) and beet plant (in red). With the original image on the left and the attention map (right) obtained with ViT-B16 model. The attention map shows the model's attention on the different plants: with a dark blue and purple colour pixel representing the attention on the weeds and a light blue colour pixel representing the beet plant.

The figure shows two side-by-side images. On the left is the 'Original image', which is a close-up of a plant patch containing green weeds and a red beet plant. On the right is the 'Attention map', which is a heatmap where different colors represent the model's attention. Dark blue and purple pixels are concentrated on the green weeds, while light blue pixels are concentrated on the red beet plant. Blue arrows point from the attention map back to the corresponding regions in the original image.

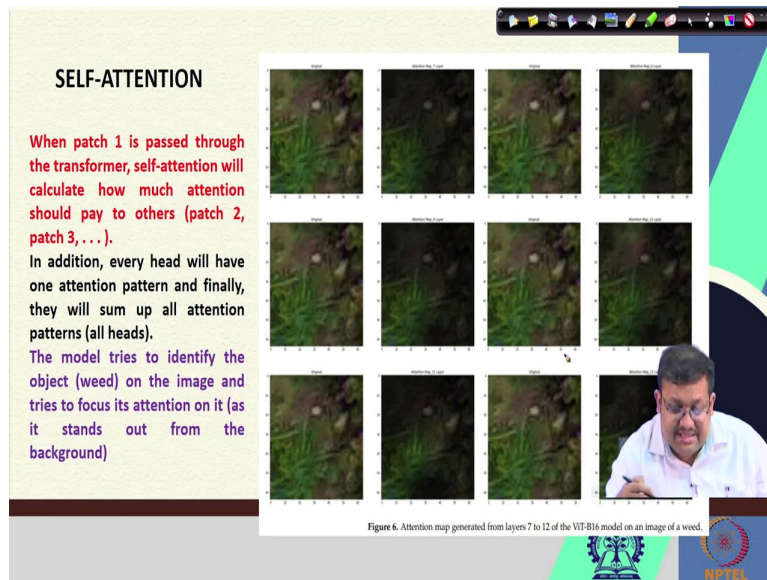
So, in this study also you can see here, this is a picture which shows the self-attention from the, or self-attention created map from an original image. So, here, you can see here

this is an attention mechanism of an image patch. So, this is the original image patch, and this is the attention. So, original image, you can see, containing these greens, which are basically the weeds, and beet plant which are red. So, we can see the clear difference between the beets as well as the weeds.

So, this is the original image. So, the original image basically goes through the self-attention and this is basically the attention map which is obtained with this ViT-B16 model, which is a Vision Transformer model, and this ViT-B16 model, they have used So, you can clearly see that this attention map shows the model attention on the different planes.

So, here, you can see some dark blue patches. So, this dark blue basically represents the attention towards the weeds, as you can see here, towards the weeds, whereas the purple color pixels are showing the attention towards the beet plant. So, you can clearly distinguish in this attention map, the weeds from the from the beets plant.

(Refer Slide Time: 17:47)

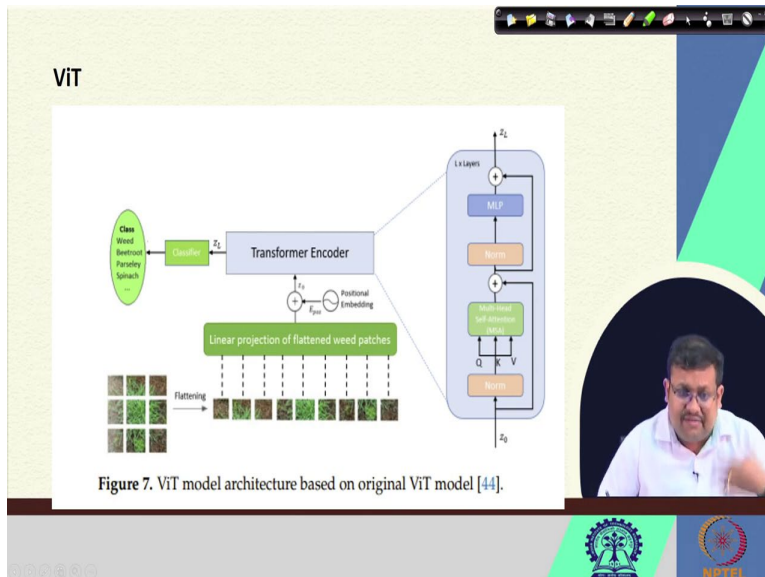


So, similar type of things, So, here you can see when Patch 1 is passed through the transformer, basically self-attention will calculate how much attention should pay to the others, like Patch 2, Patch 3, Patch 4, Patch 5, up to Patch 9 is there, suppose there are, we have divided an image into nine patches. So, when you fit that Patch 1 in the transformer, it will calculate the self-attention and it will identify how much attention to be paid to the other patches from Patch 2 to Patch 9.

So, in addition, every head will have one attention pattern and finally, they will sum up all the attention patterns, or all the heads. So, the model tries to identify the object or weed on the image and tries to focus its attention on it. So, here, you can see this is an attention map generated from layer 7 to 12 of this ViT-B16 model of an image, of a weed. So, you can see there are different original images like 7, 8, 9, 10, 11, 12, so, from the 7th layer to 12th layer patches, and they are corresponding attention maps are basically generated.

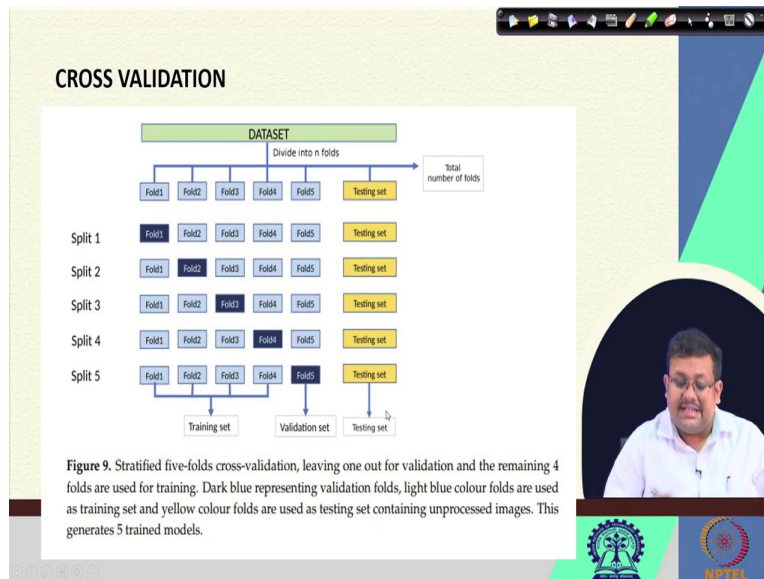
So, these attention maps calculates how much attention is to be given for these individual patches for identifying the object. In our case, this is weed. So, this is how this self-attention works.

(Refer Slide Time: 19:31)



So, once this process works, So, of course this is a construction or the ViT model architecture which is based on this, which was employed in this study. So, I have already discussed this ViT architecture previously. So, this ViT architecture was used after the self-attention.

(Refer Slide Time: 19:52)

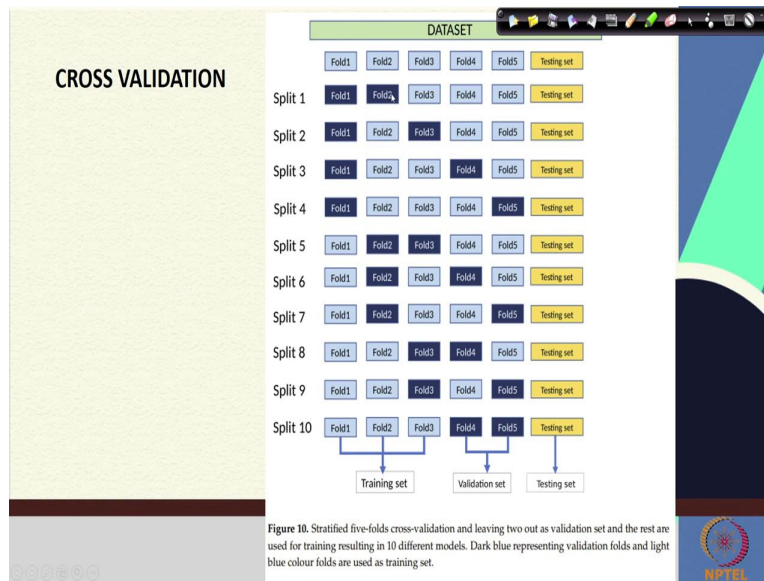


And then, finally this cross validation was used. So, there are three different types of cross validation, they have used. So, the whole dataset, as you can see, the whole dataset, they have used, they have divided into testing phase and the rest of the data, they have used in five folds. And you can see here, they are performing the K fold cross validation.

So, five-fold cross validation, they have used. So, in the split, they have, So, they have used this Fold1, Fold2, Fold3, Fold4, Fold5. So, this is the stratified five-fold cross validation. So, here they have left one fold for validation, and then they took the rest four for the training. So, you can see this dark blue representing these validation folds.

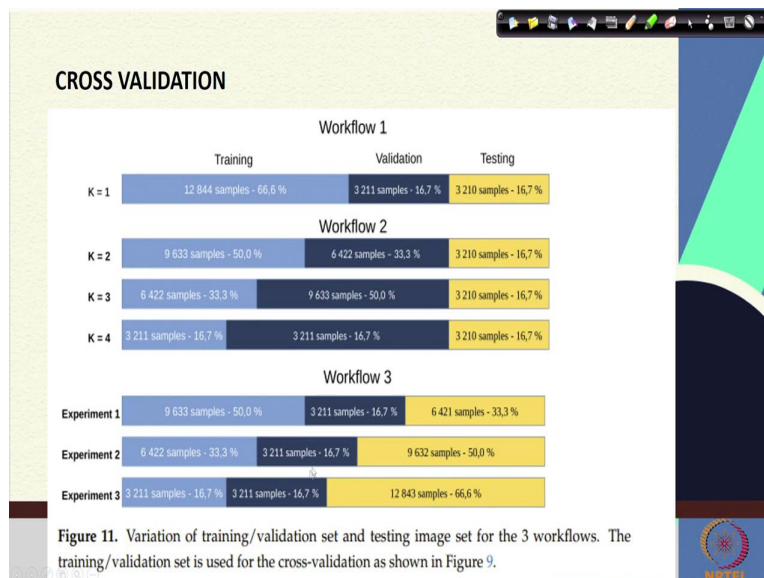
So, in the first go, they will take this Fold1 for validation. in the second go, they will take this Fold2 for validation, and they will go for the training with the rest of the four folds. So, So, on they will repeat these steps for the five times, and they will create the model and each of these steps they will test the model using the testing set. So, this is the first strategy they have used.

(Refer Slide Time: 21:15)



In the second strategy, they have used a stratified five-fold cross validation and leaving two out as validation set, and rest are used for training, resulting, ultimately, in 10 different combinations. So, here, you can see, basically that is same like the previous step, but here, instead of keeping one out, they are keeping two out. So, ultimately in this fashion they are getting 10 different combinations for training and validation and testing. So, they have used this.

(Refer Slide Time: 21:43)



And in the third workflow, they have done, again, they have kept this 9,633 samples for training and 3,211 samples for validation and 6,421 samples for testing. In the second go, they have increased the number of testing images and decreased the number of training images.

So, from 9,633 samples images in the training set, they have decreased it to 6,422 samples, and increased the testing samples from 6,421 to 9,663. And then in the third go, they have further increased the testing set, that is, up to 12,843 samples, and further reduced the samples in the training set, while keeping, in all these three experiment they have kept the validation set as constant. So, these are the three strategies they have used for model validation or cross validation.

(Refer Slide Time: 22:49)

EVALUATION METRICES

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1-Score} = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$$
$$\mu_{\text{F1-Score}} = \frac{\sum_{i=1}^N (\text{F1-Score}_i)}{N}$$
$$\sigma_{\text{F1-Score}} = \sqrt{\frac{\sum_{i=1}^N (\text{F1-Score}_i - \mu_{\text{F1-Score}})^2}{N}}$$

And the evaluation metrics, they have used a precision recall F1-score, and also they have calculated the mean and standard deviation of the F1-score, which is an important indicator of classification accuracy.

(Refer Slide Time: 23:05)

RESULTS

Table 2. Comparison between state-of-the-art CNN-based models and vision transformer models on agricultural image classification. The F1-Score has been calculated using Equation (10) with $N = 5$.

Model	$\mu_{F1-Score}$	μ_{Loss}
ViT B-16	0.994 ± 0.002	0.656
ViT B-32	0.992 ± 0.002	0.672
EfficientNet B0	0.987 ± 0.005	0.735
EfficientNet B1	0.989 ± 0.005	0.720
ResNet 50	0.992 ± 0.005	0.716

Although all network families obtain high accuracy and F1-Score, the classification of crops and weed images using vision transformer yields the best prediction performance.

So, if you see the results, the result clearly shows that okay, so, they have performed two different architecture of Vision Transformer, one is ViT-B16, another is ViT-B32, and they compare the results with the CNN architecture, or Convolutional Neural Network. We have already discussed what is Convolutional Neural Network. So, they have tried three different Convolutional Neural Network EfficientNet B0, then EfficientNet B1 and ResNet-50.

So, they have used this Convolutional Neural Network, and they have compared their mean of F1-score, and mean of loss. And while comparing all these results, you can clearly see that although all the network families obtain high accuracy, So, you can see that they are all are getting very high accuracy in terms of mean of F1-score as well as loss, but you can see the classification of crops and weed images using Vision Transformer yields the best prediction performance.

So, this Vision Transformer approach was used in this study for better classification of the weeds in the images of the crop fields. So, using this Vision Transformer, that is, splitting the images into patches, and then linear embedding of those images, and then encoding them, and then training them through self-attention have tremendous effect on the image classification accuracy, and we can clearly see that when we employ this type of strategy, that can enhance the model accuracy.

So, using this type of approach, it is now possible to improve the UAV based or in that sense, using any type of images, using any type of image sensor, you can utilize this type of Vision Transformer approach for getting better classification accuracy.

(Refer Slide Time: 25:25)

REFERENCES

- Reedha, R.; Derioquebourg, E.; Canals, R.; Hafane, A. Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. Remote Sens. 2022, 14, 592. <https://doi.org/10.3390/rs14030592>
- <https://viso.ai/deep-learning/vision-transformer-vit/>

RESULTS

Table 2. Comparison between state-of-the-art CNN-based models and vision transformer models on agricultural image classification. The F1-Score has been calculated using Equation (10) with $\mathcal{N} = 5$.

Model	$\mu_{F1-Score}$	μ_{Loss}
VIT B-16	0.994 ± 0.002	0.656
VIT B-32	0.992 ± 0.002	0.672
EfficientNet B0	0.987 ± 0.005	0.735
EfficientNet B1	0.989 ± 0.005	0.720
ResNet 50	0.992 ± 0.005	0.716

Although all network families obtain high accuracy and F1-Score, the classification of crops and weed images using vision transformer yields the best prediction performance.

So, guys, these are the two references. Reedha et al, and it was published just in this year, 2022, in the remote sensing, which is an MDPI journal. I have put the link here also. This is open access, So, you can click on this link and get it. And I would request you to again, go through this paper in detail, and then you will have more information, and you will feel more enriched while reading that paper because I have discussed the basics of it.

But when you read the whole paper it will be more clear to you how to, how they have used step by step in detailed pattern, and then you will come to know that how this type of novel strategies can be utilized in both the crop as well as, you can also devise your own methodology for soil classification also using this type of Vision Transformer approach.

So, guys, I hope you this lecture series of this week 8 was informative to you. And we have tried to answer, we have tried to pinpoint some of the contemporary advancement in deep learning domain. We have discussed the Recurrent Neural Network, we have discussed in details about the Vision Transformer approach, self-attention approach, and we have seen one example where they are being used for better classification of the image taken by the UAV.

And of course, you can utilize the same strategy into multiple domain of agriculture, in multiple objects in any domain of agriculture to classify them. And once you do that, then only you will feel more and more confident enough. So, guys, I hope that these lectures were very much informative for you, and you have learnt something new.

Feel free to let me know if you have any questions, and I will be more than happy to answer your queries. I would request you to, when you write, when you ask a question in the forum, you should be, you should ask the specific question, so, that it will be easy for us to answer that query. So, feel free to ask any questions and we will be happy to answer your queries. So, let us wrap up this week, and let us meet in our next week of lectures. Thank you.