STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, January 2026, pp 550–570.
Published online in International Academic Press (www.IAPress.org)

Effect of Preprocessing on Modelling Soil Images Captured Using Smartphone

Yudi Agusta*, Ni Komang Sri Julyantari

Informatics and Computer Faculty, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia

Abstract Knowing soil characteristics is one crucial step in the agricultural process. Soil characteristics such as NPK and pH values could differ the production quantity and quality of a farm. To know soil characteristics, various methods could be implemented including the use of tools such as Soil Test Kit (STK), and Rapid Soil Testing (RST), among others. For an extreme case, soil laboratory work is sometimes conducted. However, such a process is considered taking time and expensive to realize. Nowadays, the use of smartphones is getting common. Smartphones can capture images, in this case soil images, in no time. However, recognizing soil characteristics based on images needs more processes. Various artificial intelligence (AI) methods exist and could be used for the purpose, including artificial neural networks, convolutional neural networks, random forest, and gradient boosting, among others. This paper tries to experiment how the soil images captured using smartphone could be used to predict soil characteristics. Various image augmentation and preprocessing methods are chosen to produce images to see the effect of the two processes in the modelling. Modelling using deep learning were also conducted with an added process of using transfer learning. The results show that SwinModel, a type of transformer deep learning, performed the best compared to other methods with lower values of evaluation metrics. Gradient Boosting and Random Forest were also recommended with relatively low values of evaluation metrics. Based on the experiment results, preprocessing and augmentation were proven to enhance the quality of modelling. For preprocessing, median and homomorphic filters are recommended, and for augmentation, color, physical, and rotation-based augmentation are recommended. As the soil images used in this study were captured in a process that can easily be imitated by farmers, further implementation in the form of mobile application shows a good prospect.

Keywords Soil Images Modelling, Image Capturing Using Smartphone, Image Preprocessing and Augmentation, Transfer Learning

DOI: 10.19139/soic-2310-5070-2198

1. Introduction

Soil processing is one essential process in quality farming. Soil type measured by its fertility and acidity influences how the soil should be utilized and maintained [1, 2]. Soil conditions will also affect the way the farmer chooses the plant and fertilizers needed [3, 4, 5, 6]. The plant choice must do a check on the soil type from the beginning when selecting the type of seeds to be planted. As for fertilizers, the amount, and the timing when the fertilizers are applied, are important. The number of fertilizers applied will depend on the type of soil. Lacking or abundance of fertilizers applied could cause problems including the attack of insects, pests, and diseases [1, 7, 8, 9]. Crop weeds could also become a problem when soil management is poorly conducted. So, the information about soil types and their characteristics is important, even before the planting is started.

*Correspondence to: Yudi Agusta (Email: yudi@stikom-bali.ac.id). Informatics and Computer Faculty, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1.1. Soil Characteristics Identification

Identifying soil characteristics can be performed using various processes including the uses of Soil Test Kit (STK), Rapid Soil Testing (RST), Soil Testing Capsules (STC), Near-Infrared Spectroscopy, Spectrophotometric, and Lost on Ignition (LOI), among others [10, 11, 12, 13, 14, 15, 16]. Laboratory work by collecting soil samples and conducting experiments could also be conducted for parameters collection. Experiments are conducted to find basic soil properties by taking soil samples, drying them, and using spectrophotometer or spectrometer to provide colour values [17]. However, these approaches are known to be expensive from the farmers' point of view.

1.2. Capturing Using Smartphone

On the other hand, taking pictures using smartphones for soil analysis could face problems such as the inclusion of unnecessary and irrelevant objects. Smartphone could also experience lack of focus, so that pictures could also be exposed to blurriness. Lighting could also become a problem, as some parts of the picture contain colors that do not represent the soil condition and are affected by the lighting during the image capturing. Certain preprocessing methods to manage these conditions need to be implemented.

1.3. Soil Image Based NPK and pH Identification

1.4. Contributions

This paper tries to find the effect of image preprocessing in modelling soil images captured using a smartphone. Two analyses are conducted. Firstly, exploratory data analysis (EDA) and the grouping of soil are conducted to look at the condition of soil image data, and secondly, the classification modelling to predict NPK and pH values existing in soil with diverse setting of preprocessing, augmentation, transfer learning, grid search on modelling parameters, and smartphone types.

1.5. Organization

This paper is arranged as follows: Section 1 elaborates on the general background of the study, Section 2 describes the research methodology and materials required for the study including clustering and classification methods implemented in this study, Section 3 provides the modelling results for both clustering and classification processes, Section 4 discusses the modelling results and provides recommendations on the identification of soil characteristics using soil image data, and Section 5 concludes the report with some conclusions.

2. Methods, Materials, and Theories

2.1. Research Methodology

Research conducted following the methodology as shown in Figure 1. Research started with soil image data collection using a smartphone from agricultural fields. Collection is also conducted for soil characteristics information including NPK and pH values. Various preprocessing and augmentation methods are applied to see methods that are suitable to be used for soil image classification. Before processing, the data are explored to see the characteristics of data and the potential cluster and variations exist in the dataset. EDA is conducted in terms of the ranges, means, standard deviation, and normality of collected NPK and pH values. The clusters exploration is conducted using the Minimum Message Length (MML) image clustering/mixture modelling method with the probability bit-costings method used as evaluation.

For image preprocessing, methods considered include Remove Unwanted Elements, Replacing Outliers, Gaussian Blur, Median Filter, Bilateral Filter, White Balance, MSRCR, and Homomorphic filters. Augmentation methods investigated include Color, Physical, Rotate, and Moisture Augmentations. Selection of batch sizes, epoch, neurons, and optimizers is also conducted for ANN and CNN. For CNN, transfer learning methods are applied which include the uses of ResNet50, EfficientNetB4, ViTModel, and SwinModel. The modellings are conducted using four methods Gradient Boosting, Random Forest, ANN, and CNN.

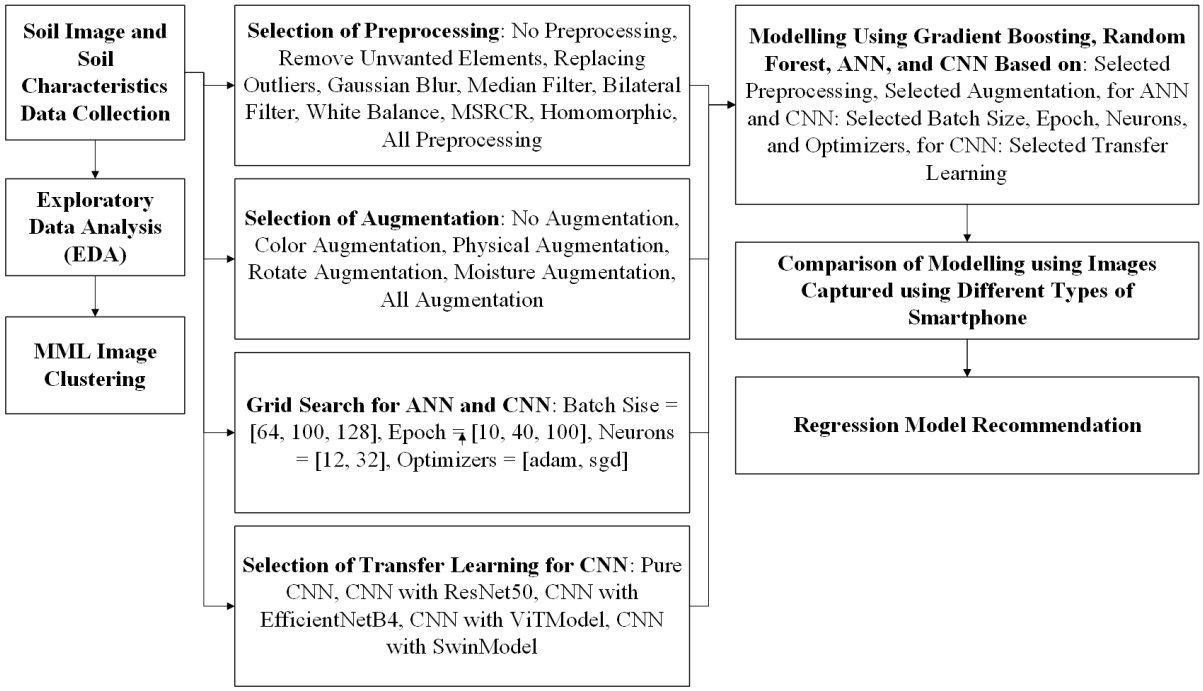

Figure 1. Research Methodology

Once the preprocessing, augmentation, batch size, epoch, neurons, optimizers, and transfer learning are evaluated, suitable methods are chosen among them to be used for soil image classification using the four methods. The comparison between modelling using images captured using several types of smartphones is also conducted. Among the selected ones, one best model will be chosen to be recommended as the best model for modelling soil images.

Some notes are included in the study. Feature extraction is conducted by implementing k-means segmentation method, selection of the most dominant class, finding the closest value to Munsell Soil Color database, normalization of image pixels, and conversion of image values into hue (h), saturation (S), and brightness (V) values, which are indicative of soil properties. This process only applies to Gradient Boosting, Random Forest, and ANN. CNN has a specific process for obtaining colour information utilizing the convolutional and max pooling layers. All four methods are evaluated using three evaluation metrics i.e., Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R2).

2.2. Soil Image Collection Using Smartphones

With ease in using smartphones to capture images and the availability of the tools on the farmer side, taking soil images using smartphones is recommended in analyzing soil characteristics. In using smartphone, some parameters need to be considered during the processes including the resolution of the resulted image and the lighting required to get the image [18]. In implementing digital photography, it is important to know that digital colour information is obtained at visible wavelengths [19] or visible/near-infrared spectroscopy [20, 21].

In using a smartphone for capturing soil images, it was also found that the best time to capture images is between 10 am to 4 pm with the lighting condition are within the values of 3500 to 70000 lux [22]. Each device also has its own characteristics. For one specific image capturing purpose, one device could produce better results compared to the others. In capturing images, several pieces of information can be obtained from each image including latitude and longitude, date and time, and altitude. These are available if GPS is set to on.

Capturing images using smartphones could have some drawbacks. These include the inclusion of irrelevant objects in images, lacking colour information affected by lack of or abundance of lighting, and blurriness when the focus is not obtained during the image capturing. These conditions require various preprocessing to make sure that the images represent the soil conditions as close as possible.

2.3. MML Mixture Modelling

Minimum Message Length (MML) mixture modelling is a type of clustering method which models data into groups. This method differs from other clustering methods in that the resulting model consists not only of the probability of the data given the model but also the prior probability of the underlying model. MML mixture modelling consists of coding two parts of the message which includes coding the number of classes, coding the relative abundance of each class, coding the model of each class, and coding the probability of the data belonging to each class [23, 24, 25, 26, 27].

Based on this definition, as also implemented in the MML estimation of statistical distributions, the estimation of the parameters of each class and the selection of the best mixed model for the data are conducted simultaneously using the same concept. Thus, MML mixture modelling has been shown to be theoretically consistent compared to other model selection methods. For the distribution used as a model for each population in mixture modelling, model selection is also conducted using the MML concept. The distribution model that has the least message length will be selected as the class distribution model.

2.4. Image Preprocessing

Images captured from the field using a smartphone could contain various issues such as blurriness, poor lighting, or obstructions. These imperfections could adversely affect the performance of modelling by introducing noises and irrelevant information. For this, a solution in the form of image cleaning is needed. Image cleaning involves removing any parts of the images that are of inadequate quality or do not meet predefined criteria or preprocessing the images by removing irrelevant objects from images. This step ensures that only high quality and relevant images are used for the modelling.

There are various methods that could be used for image cleaning. They include quality check and removal of poor-quality images. Image quality check inspects images for clarity, lighting, or presence of obstructions. This process could be done manually or automatically. Manual process involves human evaluation of images, which can be time-consuming, but ensures high accuracy.

Automatic quality check uses algorithms for blurriness detection, poor contrast, outliers' removal, median filtering, and other quality issues [10, 28]. Their quality checking could be used to improve the quality of images obtained. If the qualities of images are off the standard set, even after the image cleaning process, then the images should be removed. Removal of poor-quality images, on the other hand, discards images that are blurry, poorly lit, or contain irrelevant objects. However, the number of images could decrease significantly when this process is implemented.

Various methods could be implemented to automatically improve quality of soil images such as Median and Gaussian Filters [29] or Noise Reduction and Feature Enhancement, Gaussian Blur [30] for noise reduction and image smoothing, Bilateral Filter [31] for noise reduction without losing soil critical structure features, White Balance [32] which is crucial for soil image analysis, as it affects applications on organic matter estimation, MSRCR (Multiscale Retinex with Color Restoration) [33] which combines multi-scale illumination correction with color preservation, and Homomorphic filter [34] for enhancing soil images by separating and independently processing illumination and reflectance components.

2.5. Image Augmentation

To overcome lack of images numbers, image augmentation could also be applied for the analysis [35]. Image could be augmented based on colour, physical, rotation, and soil moisture. Color augmentation [36] is a crucial process for improving deep learning model robustness, as it provides possible variations by implementing color-based modification, manipulation, and correction. Physical augmentation [37] enhances soil image datasets by simulating

real-world physical variations. Rotation augmentation [38] could be used for improving model generalization and addressing data scarcity by performing various rotations. Moisture augmentation [39] adds moisture effects such as light reflection and pore ware retention on soil images varying moisture in available soil images.

By performing image augmentation, variation of images could be obtained, so that training could be conducted with higher accuracy.

2.6. Classification Methods

2.6.1. Artificial Neural Networks Artificial Neural Networks (ANN) is a method representing how the human brain works [12, 14]. The inputs from external are the stimulant to the inner hidden layers existing in the model to produce the expected outputs. Hidden layers consist of nodes/neurons that are connected to one or more inputs or nodes/neurons from a previously hidden layer. The number of nodes/neurons could vary and could affect the resulting model [40]. To process the input in producing an output, an activation function is installed. Various activation functions are available including Relu, tanh, and Sigmoid activation functions, among others. Sometimes, a bias value is added as part of the node processing.

In training ANN, the process could be set based on batch for a number of epochs. The number of batches could also interfere with the modelling process in terms of speed, accuracy, and generalizations [41]. The number of epochs affects the resulting model of soil analysis [42] and can have assorted options including fixed numbers, early stopping, and cyclic epochs. Optimizers are also involved in ANN modelling training, with diverse options available including SGD, and AdamW. AdamW has been proven to be superior for some cases [43].

2.6.2. Convolution Neural Networks Convolutional Neural Networks (CNN) is a type of artificial neural network with steps to preprocess the data so that the data come into the neural networks are compact enough and represent the condition better compared to the original data [44, 45, 46]. The process is conducted using filter optimization known as cascaded convolution kernels. Another layer which is also installed in the concept is the pooling layer. Pooling layers are installed to the neural networks concept so that original data are abstracted to a feature map which is commonly called activation map which reduces the dimension of data to become smaller following the size of the activation map.

In CNN, the uses of transfer learning are also introduced using distinct options including the uses in ResNet50, EfficientNetB4, Vision Transformer (ViT), and SwinModel. ResNet50 [47] is a widely used pre-trained CNN architecture for transfer learning due to its deep residual connections, robust feature extraction, and efficiency in training. EfficientNetB4 [48] is a lightweight and powerful CNN for transfer learning due to its compound scaling and computational efficiency. ViT model [49] is used as pre-trained transfer learning which is based on large datasets. Swin Transformer [50] bridges the gap between ViT model and CNN with local-window self-attention, shifted-window partitioning, and hierarchical feature maps.

2.6.3. Random Forest Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees [51]. This method can manage mixed data types including numerical and categorical data. This makes it suitable for datasets that include diverse types of features, which as numerical soil measurements and weather conditions. This method also introduces robustness and the ability to avoid overfitting, which is particularly beneficial when dealing with noise or incomplete data. This modelling output of this method is also interpretable which could provide insights into feature importance, helping common users to understand which factors are the most influential in the resulting prediction model.

2.6.4. Gradient Boosting Gradient Boosting is also an ensemble learning method that builds models sequentially, with each new model correcting errors made by the previous processes [52]. This method has the advantages of high accuracy results where this method often outperforms other methods in terms of prediction accuracy. This is achieved due to its iterative nature and focus on error reduction. This method can also capture complex patterns and can model intricate relationships between features and targets, making it suitable for complex datasets. This method is also customizable, which provides many hyperparameters that can be tuned to optimize model performance for specific datasets.

2.7. Model Validation

For modelling validation, probability bit-costings is used for mixture modelling, and root means squared error (RMSE), mean absolute error (MAE), and R-squared (R^2) are used for regression.

2.7.1. Probability bit-costings Probability bit-costing method is a validation method used for clustering models. The method calculates the negative logarithm of the probability of test data given the model obtained from the training data. The probability bit-costings of the data given the model is given by:

$$probability\ bit-costings = -\log P(x_{test}|f(\mu,\sigma|x_{train})) \tag{1}$$

where *P*() is the probability value, x_{test} the testing data, f () is the likelihood function, μ is the means, σ is the standard deviation, and x_{train} is the training data.

2.7.2. Root Means Squared Error (RMSE) RMSE is commonly used for evaluating the modelling results and looking at how good the prediction results are using the resulting models compared to the original targets. The root means squared error (RMSE) is calculated using the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2} \tag{2}$$

where n the number of data, Y_i the observed data, and $\widehat{Y}_i$ the predicted data.

2.7.3. Mean Absolute Error (MAE) Mean absolute error (MAE) the average of the absolute differences between the predicted values and the actual values. MAE is calculated using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) \tag{3}$$

where n the number of data, Y_i is the observed data, and $\widehat{Y}_i$ the predicted data.

2.7.4. R-Squared R-squared (R^2) is commonly used to assess the good fitness of a regression model. R^2 is calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \acute{Y}_i)^2} \tag{4}$$

where n the number of data, Y_i the observed data, $\widehat{Y}_i$ the predicted data, and $\acute{Y}_i$ the means of the actual value.

3. Research Results

3.1. Data Preparation

Image data were collected from a number of agricultural fields sites in Bali Island, Indonesia using a smartphone along with the collection of other soil characteristics such as NPK and pH values. The types of smartphones used in this study were POCO M4 PRO with Android operating systems supported by a camera specification of 64 MP f/1.8, and iPhone 13 with IOS operating systems supported by a camera specification of dual camera 12 MP. The soil images were taken from 10 – 20 cm above the ground, and before the image capturing, the lands have been cleaned up to a certain degree from irrelevant objects. The image capturing processes were conducted during the sunny days or partly cloudy days, between 10:00 am to 14:00 pm. Total images collected were 124 soil images. Samples of soil images and their NPK and pH values are provided in Table 1.

Table 1. Samples of Soil Images

				
NPK=6.1, pH=7.0	NPK=6.0, pH=7.0	NPK=5.3, pH=6.7	NPK=5.2, pH=6.8	NPK=5.5, pH=6.5
				
NPK=6.0, pH=7.0	NPK=4.1, pH=6.5	NPK=5.0, pH=6.7	NPK=5.0, pH=6.9	NPK=5.0, pH=7.0

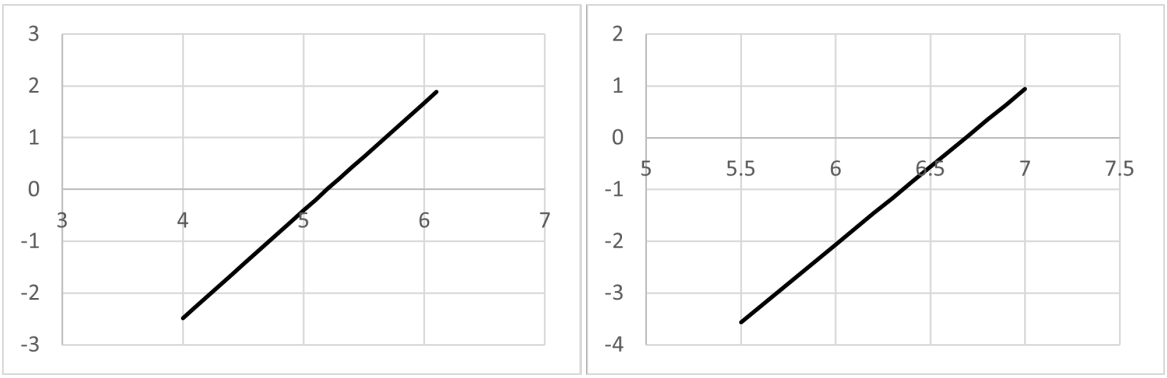

Figure 2. QQPlots of NPK and pH Values

For NPK and pH values data collection, we use a tool called Doctor Plant to obtain the values directly from the soil. The ranges of NPK and pH values obtained were between 4.0 – 6.1 and 5.5 – 7.0, respectively. The means and standard deviation of NPK and pH values are 5.195 (± 0.481) and 6.588 (± 0.333). The QQPlots of NPK and pH values are as in Figure 2 (left) and Figure 2 (right). Both QQPlots showed straight lines, which proved that the obtained values were normally distributed.

During modeling, the soil images will undergo preprocessing, augmentation, and feature extraction. The implementation of these processes will vary for each method and adapt to the characteristics of the method. The feature extraction involves implementing k-means segmentation method, selection of the most dominant class, finding the closest value to Munsell Soil Color database, normalization of image pixels, and conversion of image values into hue (h), saturation (S), and brightness (V) values.

3.2. Data Clustering of Soil Image Data

Data clustering is conducted to see the variations existing in soil image data and make sure that the source data is not coming form one pattern. Data clustering was conducted using the MML mixture modelling method. The method can naturally obtain the most appropriate number of clusters existing in the dataset by simultaneously performing parameter estimation and mixture model selection. In the modelling, each cluster was assumed to come from a Gaussian distribution. Data used for modelling is the original data without augmentation. Three cases of data

preprocessing are investigated. They are data without preprocessing, data with selected preprocessing, and data with complete preprocessing. For complete preprocessing, the methods applied were remove unwanted elements, replace outliers, gaussian blur, median filter, and bilateral filter. For selected preprocessing, the methods applied include removing unwanted elements. To check which approach performed better, the probability bit-costings calculation is applied, where the original data is divided into training and test data with a composition of 80:20. The results are provided in Table 2.

Table 2. Resulting MML Mixture Models for Data Without, With Selected and Complete Preprocessed Data

Indicators	No Preprocessing	Selected Preprocessing	Complete Preprocessing
(1)	(2)	(3)	(4)
No of Classes	7	5	5
Probability Bit Costing	48,608	46,329	44,412

Based on the results, it is shown that modelling with complete preprocessing has resulted in the least probability bit-costing with five clusters. This means that the clustering with complete preprocessing produced less variations between the resulting models with the testing data.

Modelling all data using the modelling which has the lowest probability bit-costings, i.e., modelling with complete preprocessing, resulted in an 8 (eight) class model with the characteristics of each class, represented by mixing proportion, means and standard deviation of each variable, are as in Table 3.

Table 3. Characteristics of Groups Resulted from MML Mixture Modelling for Data with a Complete Preprocessing

Indicators	Class 1			Class 2		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mixing Proportion	19,817			9,690		
Variable	h	V	S	h	V	S
Means	2,692	3,813	14,637	5,935	6,264	17,280
Standard Deviation	0,029	2,217	2,020	1,408	0,511	1,145
	Class 3			Class 4		
Mixing Proportion	6,667			6,667		
Means	2,930	4,000	9,000	10,215	5,500	12,000
Standard Deviation	0,001	0,001	4,243	0,049	0,707	0,001
	Class 5			Class 6		
Mixing Proportion	11,997			17,160		
Means	5,180	2,722	12,756	7,769	3,520	13,355
Standard Deviation	0,001	0,520	1,605	0,115	1,806	7,687
	Class 7			Class 8		
Mixing Proportion	18,003			10,000		
Means	6,691	3,000	8,904	5,347	5,000	10,000
Standard Deviation	3,097	0,001	1,720	2,626	0,001	0,001

As shown in Table 2, the data were naturally grouped into eight classes with mixing proportions of each class are 19,817, 9,690, 6,667, 6,667, 11,997, 17,16, 18,003, and 10,0, respectively. The characteristics of each class for the eight resulting groups are quite different from one to another. This shows that there are enough variations in the soil images dataset.

3.3. Classification Model

3.3.1. Data Preprocessing To observe the effect of preprocessing on soil image modeling, several image preprocessing methods were studied. Their resulting accuracies were observed when used in the modeling process.

The methods studied included Bilateral Filter, Gaussian Blur, Homomorphic filter, Median filter, and MSRCR. The modeling results were also compared with modeling without preprocessing and modeling with all preprocessing methods applied.

The four modeling methods, i.e., ANN, CNN, Gradient Boosting, and Random Forest, were studied. The setting for ANN, CNN, Gradient Boosting, and Random Forest modelling are as in Table 4. The dataset is divided into 80:20 of training:testing data.

Table 4. Parameter Settings for Each Investigated Method

Methods	Parameter Setting
(1)	(2)
ANN	Two hidden layers of 12 and 6 neurons, activation function = sigmoid, use bias, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 100
CNN	Two Convolutional layers (3x3), two Max Pooling layers (2x2), activation function = relu, ANN with two hidden layers of 12 and 6 neurons, activation function = sigmoid, use bias, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 32
Gradient Boosting	Estimator number = 100, learning rate = 0.01, max_depth = 20, loss function = squared_error
Random Forest	Estimator number = 100

Table 5. The resulting accuracies of modelling with various preprocessing methods

Classification Methods	Preprocessing	RMSE		MAE		R^2	
		NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ANN	All Processing	0.2215	**0.3105**	0.1840	**0.2680**	-6.7793	**-144.3200**
	Bilateral Filter	0.2131	0.3745	0.1624	0.3212	-6.1988	-210.4264
	Gaussian Blur	0.2381	0.3354	0.2078	0.3142	-7.9875	-168.5684
	Homomorphic	**0.1850**	**0.3126**	**0.1565**	0.2955	**-4.4272**	-146.2925
	Median Filter	0.2631	0.3828	0.2105	0.3642	-9.9709	-219.8805
	MSRCR	0.2274	0.3375	0.1975	0.2984	-7.1957	-170.7168
	Removing Unwanted	0.2069	0.3679	0.1771	0.3429	-5.7879	-203.0346
	Replacing Outliers	0.2064	0.3421	0.1816	0.3238	-5.7537	-175.4878
	White Balance	0.2702	0.2702	0.2283	**0.2547**	-10.5717	**-109.0700**
	No Preprocessing	**0.1729**	0.3136	**0.1438**	0.2928	**-3.7373**	-147.2446
CNN	All Processing	0.1779	0.2947	0.1475	0.2369	-4.0152	-129.938
	Bilateral Filter	0.2177	**0.2616**	0.1845	**0.2252**	-6.5163	**-102.2036**
	Gaussian Blur	0.2191	0.2812	0.1606	0.2608	-6.6081	-118.2249
	Homomorphic	0.1885	0.3280	0.1589	0.2922	-4.6310	-161.1660
	Median Filter	**0.1490**	0.2756	**0.1273**	0.2463	**-2.5214**	-113.4776
	MSRCR	0.1776	0.3227	0.1407	0.2707	-4.0010	-155.9831
	Removing Unwanted	0.2117	0.2882	0.1765	0.2437	-6.1036	**124.2292**
	Replacing Outliers	**0.1331**	0.2720	**0.1065**	**0.2273**	**-1.8064**	-110.5842
	White Balance	0.1973	**0.2709**	0.1579	0.2385	-5.1703	-109.6827
	No Preprocessing	0.1748	0.3322	0.1518	0.3129	-3.8459	-165.3680
	All Processing	0.0807	0.0259	0.0574	0.0230	-0.0326	-0.0085
	Bilateral Filter	0.1048	0.0455	0.0835	0.0414	-0.7424	-2.1279
	Gaussian Blur	**0.0770**	0.0408	**0.0558**	0.0350	**0.0599**	-1.5144

Classification Methods	Preprocessing	RMSE		MAE		R^2	
		NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gradient Boosting	Homomorphic	**0.0800**	0.0244	0.0584	0.0232	**-0.0143**	**0.1013**
	Median Filter	0.0980	**0.0213**	0.0795	**0.0177**	-0.5225	**0.3146**
	MSRCR	0.0807	0.0259	0.0574	0.0230	-0.0326	-0.0085
	Removing Unwanted	0.0842	0.0444	0.0600	0.0387	-0.1229	-1.9747
	Replacing Outliers	0.1017	0.0388	0.0777	0.0301	-0.6404	-1.2649
	White Balance	0.0807	**0.0259**	**0.0574**	**0.0230**	-0.0326	-0.0085
	No Preprocessing	0.1126	0.0362	0.0891	0.0276	-1.0105	-0.9775
Random Forest	All Processing	0.0808	0.0259	0.0574	0.2680	-0.0344	-0.0123
	Bilateral Filter	0.1155	0.0430	0.0887	0.3212	-1.1158	-1.7886
	Gaussian Blur	0.0909	0.0370	0.0628	**0.0304**	-0.3087	-1.0696
	Homomorphic	**0.0797**	**0.0256**	0.0676	0.2955	**-0.0069**	**0.0111**
	Median Filter	0.0858	**0.0192**	0.0723	**0.0166**	-0.1678	**0.4433**
	MSRCR	0.0808	0.0259	**0.0574**	0.2984	-0.0344	-0.0123
	Removing Unwanted	0.0863	0.0375	0.0581	0.0342	-0.1806	-1.1179
	Replacing Outliers	0.1003	0.0415	0.0685	0.0308	-0.5948	-1.5919
	White Balance	**0.0808**	0.0259	**0.0574**	0.2547	**-0.0344**	-0.0123
	No Preprocessing	0.1171	0.0405	0.0877	0.0332	-1.1738	-1.4678

Judging from the modeling results in Table 5, it can be seen that various preprocessing methods have varying effects on the modeling methods used. Modeling for predicting pH and NPK values also showed different trends. For soil NPK modeling, the Homomorphic Filter performed well for both the ANN and Random Forest methods. For ANN, modeling without preprocessing also showed satisfactory results. For CNN, the Median Filter and Replacing Outliers had a significant effect on modeling. Meanwhile, for Gradient Boosting, Gaussian Blur had a moderate effect on modeling. Lastly, for Random Forest, White Balance also showed reliable results.

For soil pH modeling, modeling with preprocessing using all methods and modeling using White Balance for preprocessing had the best accuracy for the ANN method. For CNN, the Bilateral Filter showed better results. Meanwhile, for Gradient Boosting, the Median Filter produced better results. Lastly, like Gradient Boosting, the Median Filter also showed superior results in modeling using the Random Forest method.

3.3.2. Data Augmentation To observe the effect of augmentation on soil image modeling, several augmentation methods were studied. Their resulting accuracies were observed when used in the modeling process. The augmentation methods studied included color augmentation, moisture augmentation, physical augmentation, and rotation augmentation. The modeling results were also compared with modeling without augmentation and modeling with all augmentation methods applied. Here, the four modeling methods, i.e., ANN, CNN, Gradient Boosting, and Random Forest, were also studied. Parameter settings used for the four methods are set the same as the experiments reported in Section 3.3.1.

Table 6. The resulting accuracies of modelling with various augmentation methods

Classsification Methods	Augmentation	RMSE		MAE		R^2	
		NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ANN	All Augmentation	0.2564	0.3500	0.2090	0.3172	-8.6308	-111.5300
	Color Augmentation	**0.1951**	**0.2796**	**0.1615**	**0.2505**	**-4.5723**	**-70.8365**
	Moisture Augmentation	0.2523	0.3155	0.2120	0.2957	-8.3210	-90.4266

Classsification Methods	Augmentation	RMSE		MAE		R^2	
		NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CNN	Physical Augmentation	**0.1822**	**0.2335**	**0.1540**	**0.2037**	**-3.8640**	**-49.1044**
	Rotate Augmentation	0.2406	0.3095	0.1789	0.2821	-7.4779	-87.0178
	No Augmentation	0.2276	0.3517	0.1846	0.3384	-7.2091	-185.465
	All Augmentation	**0.1344**	**0.1986**	0.1047	**0.1714**	**-1.6470**	**-35.2424**
	Color Augmentation	**0.1220**	0.2514	**0.0953**	0.2157	**-1.1781**	-57.0881
	Moisture Augmentation	0.1616	0.2500	0.1182	0.1810	-2.8245	-56.4274
	Physical Augmentation	0.1362	**0.1943**	**0.1040**	**0.1732**	-1.7150	**-33.6940**
	Rotate Augmentation	0.1664	0.2443	0.1266	0.2108	-3.0556	-53.8207
	No Augmentation	0.1914	0.2544	0.1643	0.2270	-4.8081	-96.567
Gradient Boosting	All Augmentation	0.0894	0.0319	0.0690	**0.0239**	-0.1713	0.0651
	Color Augmentation	0.0750	0.0358	0.0563	0.0277	0.1764	-0.1783
	Moisture Augmentation	0.0818	0.0346	0.0621	0.0290	0.0198	-0.0975
	Physical Augmentation	**0.0598**	**0.0295**	**0.0451**	**0.0212**	**0.4767**	**0.2003**
	Rotate Augmentation	**0.0744**	**0.0304**	0.0573	0.0250	**0.1901**	**0.1512**
	No Augmentation	0.0877	0.0455	**0.0527**	0.0407	-0.2184	-2.1249
Random Forest	All Augmentation	0.0982	0.0368	0.0757	0.3172	-0.4118	-0.2436
	Color Augmentation	0.0764	0.0369	0.0594	**0.0280**	0.1447	-0.2541
	Moisture Augmentation	0.0875	0.0337	0.0675	0.0291	-0.1207	-0.0445
	Physical Augmentation	**0.0667**	**0.0333**	**0.0483**	**0.0268**	**0.3490**	**-0.0193**
	Rotate Augmentation	**0.0654**	**0.0335**	**0.0533**	**0.0285**	**0.3741**	**-0.0329**
	No Augmentation	0.1000	0.0395	0.0771	0.0327	-0.5856	-1.3557

Judging from the modeling results in Table 6, it can be seen that various augmentation methods have varying effects on the modeling methods used. Modeling for predicting pH and NPK values also showed different trends. For soil NPK modeling, the color augmentation performed well for both the ANN and CNN methods. For ANN, physical augmentation also produced better results. For CNN, the modelling with all augmentation had a significant effect on modeling. Meanwhile, for Gradient Boosting and Random Forest, physical and rotation augmentations had an advantageous effect on modeling.

For soil pH modeling, similar to NPK, modeling with color and physical augmentations had the best accuracy for the ANN method. For CNN, modelling with all augmentation and physical augmentation showed better results. Meanwhile, for Gradient Boosting and Random Forest, physical and rotation augmentations had a favorable effect on modeling.

3.3.3. CNN Variations with Transfer Learning Four transfer learning methods were studied including ViTModel, SwinModel, ResNet50, and EfficientNetB4. The parameter setting of each CNN with transfer learning modelling were as in Tabel 7. ViTModel implemented google/vit-base-patch16-224 as a processor, whereas SwinModel implemented microsoft/swin-tiny-patch4-window7-224. Both ResNet50 and EfficientNetB4 used pretrained model of imagenet.

Table 8 shows the accuracy results of CNN modelling using transfer learning. Compared to basic CNN, all CNN modelling using transfer learning produced better results. Among the four, ResNet50 produced the best results for both NPK and pH. SwinModel also produced satisfactory results for NPK, whereas EfficientNetB4 resulted in reliable results for predicting pH values.

3.3.4. Grid Search for CNN and ANN As modelling ANN and CNN need setting for their training processes, grid search for ANN and CNN modelling was also conducted in terms of their batch size, epoch, neuron, and optimizers

Table 7. Parameter Settings for Each CNN Modelling with Transfer Learning

Methods	Parameter Setting
(1)	(2)
Basic CNN	Two Convolutional layers (3x3), two Max Pooling layers (2x2), activation function = relu, ANN with two hidden layers of 12 and 6 neurons, activation function = sigmoid, use bias, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 32
ViTModel CNN	ImageProcessor = google/vit-base-patch16-224, dropout = 0.1, batch size = 16, epoch = 100, optimizer = adamW
SwinModel CNN	ImageProcessor = microsoft/swin-tiny-patch4-window7-224, dropout = 0.1, batch size = 16, epoch = 100, optimizer = adamW
ResNet50 CNN	Pretrained model = imagenet, hidden layer activation function = relu, output layer activation function = sigmoid, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 100
EfficientNetB4 CNN	Pretrained model = imagenet, hidden layer activation function = relu, output layer activation function = sigmoid, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 100

Table 8. The resulting accuracies of CNN modelling with various transfer learnings

CNN Variations	RMSE		MAE		R^2	
	NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Basic CNN	0.2636	0.3036	0.219	0.2531	-10.0161	-137.937
ViTModel CNN	0.1681	0.1666	0.1303	0.1288	-3.4787	-40.8344
SwinModel CNN	**0.1151**	0.0952	**0.0921**	0.0732	**-1.1011**	-12.6756
ResNet50 CNN	**0.0873**	**0.0257**	**0.066**	**0.0225**	**-0.2083**	**0.0029**
EfficientNetB4 CNN	0.1634	**0.0302**	0.1427	**0.0214**	-3.2324	**-0.3772**

method. The search is set with the following values: batch size = [64, 100, 128], epoch = [10, 40, 100], neuron number = [12, 32], and optimizer = [adam, sgd].

Table 9. The resulting accuracies of ANN modelling with various batch size, epoch, neuron, and optimizers

Batch Size	Epoch	Neuron	Optimizer	ANN					
				RMSE		MAE		R^2	
				NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
64	10	12	adam	0.2697	0.4503	0.2383	0.4338	-10.5341	-304.6888
			sgd	0.3597	0.4443	0.3174	0.4277	-19.5082	-296.6664
		32	adam	0.3168	0.3990	0.2729	0.3696	-14.9100	-238.9684
			sgd	0.3246	0.5201	0.2837	0.4949	-15.7059	-406.8668
	40	12	adam	0.2739	0.4102	0.2342	0.3790	-10.8971	-252.7514
			sgd	0.3322	0.4154	0.2808	0.3958	-16.4997	-259.1142
		32	adam	0.2519	0.3623	0.2144	0.3413	-9.0570	-196.9279
			sgd	0.3730	0.3641	0.3453	0.3515	-21.0593	-198.9240
	100	12	adam	0.2038	0.3321	0.1751	0.3054	-5.5812	-165.3230
			sgd	0.2759	0.3094	0.2528	0.2976	-11.0657	-143.3378
		32	adam	**0.1454**	0.2328	**0.1272**	0.2149	-2.3521	-80.7433

Batch Size	Epoch	Neuron	Optimizer	ANN					
				RMSE		MAE		R^2	
				NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			sgd	0.2271	0.3280	0.2050	0.3045	-7.1736	-161.2494
100	10	12	adam	0.3633	0.4221	0.3063	0.4040	-19.9198	-267.6659
			sgd	0.3327	0.3911	0.2859	0.3640	-16.5500	-229.6360
		32	adam	0.3241	0.4510	0.2793	0.4149	-15.6496	-305.6088
			sgd	0.3092	0.5120	0.2707	0.4909	-14.1585	-394.2298
	40	12	adam	0.2443	0.3622	0.2100	0.3344	-8.4623	-196.7458
			sgd	0.2809	0.3789	0.2587	0.3573	-11.5080	-215.5042
		32	adam	0.2668	0.3955	0.2283	0.3491	-10.2854	-234.7743
			sgd	0.3923	0.4391	0.3731	0.4124	-23.4030	-289.7225
	100	12	adam	0.2141	0.3038	0.1783	0.2712	**6.2658**	-138.1109
			sgd	0.2649	0.2872	0.2189	0.2749	-10.1254	-123.3395
		32	adam	**0.1407**	**0.1567**	**0.1199**	**0.1352**	-2.1389	**-36.0186**
			sgd	0.3041	0.3586	0.2806	0.3484	-13.6631	-192.8624
128	10	12	adam	0.3521	0.4550	0.3267	0.4356	-18.6498	-311.1986
			sgd	0.3467	0.4778	0.3129	0.4565	-18.0536	-343.1317
		32	adam	0.3262	0.4329	0.2983	0.4226	**15.8694**	-281.5240
			sgd	0.3546	0.4561	0.3269	0.4254	-18.9287	**312.6675**
	40	12	adam	0.2969	0.4893	0.2405	0.4592	-12.9696	-360.0069
			sgd	0.3523	0.4239	0.3024	0.4088	-18.6813	-269.9410
		32	adam	0.2001	0.3464	0.1741	0.3248	-5.3446	-179.8705
			sgd	0.2434	0.4786	0.2230	0.4600	-8.3953	-344.4117
	100	12	adam	0.2186	0.3652	0.1673	0.3463	-6.5780	-200.1268
			sgd	0.2065	0.2857	0.1806	0.2735	-5.7576	-122.0536
		32	adam	0.2035	**0.2156**	0.1632	**0.1848**	-5.5640	-69.0839
			sgd	0.2317	0.3823	0.2052	0.3443	-7.5124	-219.3135

Table 10. The resulting accuracies of CNN modelling with various batch size, epoch, neuron, and optimizers

Batch Size	Epoch	Neuron	Optimizer	CNN					
				RMSE		MAE		R^2	
				NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
64	10	12	adam	0.2509	0.4187	0.2177	0.3895	-8.9828	-263.2876
			sgd	0.3865	0.4733	0.3516	0.4515	-22.6858	-336.6830
		32	adam	0.2780	0.4109	0.2275	0.3687	-11.2494	-253.5485
			sgd	0.3809	0.4500	0.3519	0.4245	-22.0013	-304.2992
	40	12	adam	0.1605	0.3068	0.1209	0.2851	-3.0813	-140.9132
			sgd	0.3423	0.3563	0.3137	0.3218	-17.5708	-190.3823
		32	adam	0.2370	0.2374	0.1979	0.2111	-7.9042	-83.9510
			sgd	0.3038	0.3861	0.2779	0.3574	-13.629	-223.7741
	100	12	adam	0.1863	0.3384	0.1530	0.2978	-4.5012	-171.6625
			sgd	0.2161	0.3611	0.1917	0.3381	-6.3998	-195.5482
		32	adam	**0.1320**	**0.1928**	**0.1152**	**0.1697**	**-1.7609**	-55.0532

Batch Size	Epoch	Neuron	Optimizer	CNN					
				RMSE		MAE		R^2	
				NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			sgd	0.1878	0.2176	0.1665	0.1979	-4.5889	-70.4045
100	10	12	adam	0.2439	0.4439	0.2057	0.4348	-8.4279	-296.0308
			sgd	0.3183	0.4439	0.2913	0.4010	-15.0643	-296.0630
		32	adam	0.2385	0.3147	0.1837	0.2767	-8.0189	-148.2951
			sgd	0.4043	0.3831	0.3630	0.3704	-24.9118	-220.3156
	40	12	adam	0.2636	0.3431	0.2251	0.3221	-10.0113	-176.4808
			sgd	0.2221	0.3809	0.1951	0.3506	-6.8168	-217.7854
		32	adam	0.2595	0.3420	0.1985	0.3150	-9.6756	-110.0697
			sgd	0.2415	0.2866	0.1586	0.2474	-8.2490	-175.3180
	100	12	adam	0.1993	0.2866	0.1586	0.2474	-5.2948	-122.8086
			sgd	0.1720	0.3166	0.1479	0.2972	-3.6879	-150.0910
		32	adam	0.2153	0.2073	0.1611	0.1706	-6.3518	-63.8072
			sgd	0.1588	0.2735	0.1409	0.2520	-2.9958	-111.7871
128	10	12	adam	0.3014	0.3871	0.2355	0.3776	-13.4015	-224.8900
			sgd	0.3992	0.4630	0.3669	0.4483	-24.261	-322.1641
		32	adam	0.1822	0.3104	0.1572	0.3012	-4.2631	-144.3059
			sgd	0.2969	0.4235	0.2743	0.4084	-12.9711	**269.3656**
	40	12	adam	0.2695	0.3339	0.2432	0.3155	-10.5139	-167.1312
			sgd	0.2357	0.3769	0.2119	0.3335	-7.8040	-213.2255
		32	adam	0.2115	0.2252	0.1618	0.2046	-6.0892	-75.4806
			sgd	0.2522	0.3549	0.2106	0.3365	-9.0855	-188.9204
	100	12	adam	0.2520	0.3312	0.1826	0.2991	-9.0694	-164.3378
			sgd	0.2022	0.3034	0.1789	0.2902	-5.4815	-137.7502
		32	adam	**0.1165**	**0.1865**	**0.0860**	**0.1633**	**-1.1511**	**-51.4368**
			sgd	0.2144	0.2264	0.1863	0.2063	-6.2875	-76.2794

Based on the results shown in Table 9 and Table 10, for ANN and CNN, respectively, the best results obtained when the training parameters were set to agam for the optimizers, 32 for the neuron size, and 100 for epoch size. For the batch size, for ANN, batch size of 100 produced better results, whereas for CNN, batch size of 128 performed better than the others.

3.3.5. Modelling using Selected Parameters and Methods Based on the experiments conducted previously, selected modellings were conducted with various selected settings of parameters for soil image dataset captured using two types of smartphones, i.e., POCO M4 PRO with Android operating systems and iPhone 13 with IOS operating systems. The methods further investigated were Gradient Boosting, Random Forest, ANN, EfficientNetB4 CNN, ResNet50 CNN, and SwinModel CNN. Parameter set for the modelling are as in Table 11.

Table 12 shows modelling results based on images captured using POCO M4 PRO Android. Based on the results shown in the table, SwinModel CNN has shown the best results compared to the others, with Gradient Boosting coming second. These results are the contrary of the results shown in Table 6 where ResNet50 outperformed the other CNN. The additional processes before the modelling including augmentation and preprocessing made SwinModel CNN have enough variations on the data to be used for training. EfficientNetB4 also came second for modelling pH value. This is also the same case of the availability of variations in data training.

As shown in Table 13, modelling based on images captured using iPhone 13 IOS resulted in similar patterns, where SwinModel CNN performed the best among the method investigated for both NPK and pH. Gradient

Table 11. Parameter Settings for Each Selected Method

Methods	Parameter Setting
(1)	(2)
Gradient Boosting	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], estimator number = 100, learning rate = 0.01, max_depth = 20, loss function = squared_error
Random Forest	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], estimator number = 100
ANN	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], two hidden layers of 12 and 6 neurons, activation function = sigmoid, use bias, optimizer = adam, loss function = MSE, epoch = 16, batch_size = 100
EfficientNetB4 CNN	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], EfficientNet pretrained model = imagenet, hidden layer activation function = relu, output layer activation function = sigmoid, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 16
ResNet50 CNN	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], Reset pretrained model = imagenet, hidden layer activation function = relu, output layer activation function = sigmoid, optimizer = adam, loss function = MSE, epoch = 100, batch_size = 16
SwinModel CNN	Augmentation [Color, Physical, Rotation], preprocessing = [Median, Homomorphic], ImageProcessor = microsoft/swin-tiny-patch4-window7-224, epoch = 100, batch size = 16, optimizer = adamW

Table 12. Accuracies of selected modelling with selected preprocessing and augmentation for POCO M4 PRO Android

Classification Methods	RMSE		MAE		R^2	
	NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gradient Boosting	**0.1008**	**0.0393**	**0.0838**	0.0349	**-0.0190**	**-0.0444**
Random Forest	0.1010	0.0396	0.0841	0.0351	-0.0240	-0.0584
ANN	0.1024	0.1104	0.0851	0.0891	-0.0522	-7.2523
EfficientNetB4 CNN	0.1363	0.0469	0.0985	**0.0274**	-0.8638	-0.4907
ResNet50 CNN	0.1677	0.0470	0.1357	0.0277	-1.8236	-0.4922
SwinModel CNN	**0.0596**	**0.0299**	**0.0376**	**0.0212**	**0.6439**	**0.3956**

Table 13. Accuracies of selected modelling with selected preprocessing and augmentation methods for iPhone 13 IOS

Classification	RMSE		MAE		R^2	
Methods	NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gradient Boosting	**0.05712**	**0.0372**	**0.0395**	**0.033**	**-0.01098**	**-0.02756**
Random Forest	0.05738	0.03754	0.03984	0.03328	-0.01898	-0.0461
ANN	0.05738	0.08868	0.03872	0.08302	-0.02026	-6.50766
EfficientNetB4 CNN	0.12084	0.06242	0.10926	0.05048	-4.7469	-1.89052
ResNet50 CNN	0.17088	0.06168	0.15968	0.04948	-8.04398	-1.82338
SwinModel CNN	**0.027125**	**0.022275**	**0.01745**	**0.0172**	**0.7461**	**0.631225**

Boosting also came second, which still shows lower evaluation of metric values. However, all the results of modelling using images captured using iPhone 13 IOS are more accurate in predicting NPK and pH values with

less evaluation of metric values compared to modelling based on images captured using POCO M4 PRO Android. This also shows that type of smartphone influences the accuracy in predicting the value of NPK and pH.

Table 14. The resulting accuracies of Gradient Boosting and SwinModel modelling with Setting Variations

Methods	Setting Combination	RMSE		MAE		R^2	
		NPK	pH	NPK	pH	NPK	pH
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gradient Boosting	No Preproces & Augmentation	0.1126	0.0362	0.0891	0.0276	-1.0105	-0.9775
	With Preprocessing	0.0770	0.0213	0.0558	0.0177	0.0599	0.3146
	With Augmentation	0.0598	0.0295	0.0451	0.0212	0.4767	0.2003
	With Both (POCO M4 PRO)	0.1008	0.0393	0.0838	0.0349	-0.0190	-0.0444
	With Both ()iPhone 13)	0.0571	0.0372	0.0395	0.0330	-0.0110	-0.0276
SwinModel	No Preproces & Augmentation	0.1151	0.0952	0.0921	0.0732	-1.1011	-12.6756
	With Both (POCO M4 PRO)	0.0596	0.0299	0.0376	0.0212	0.6439	0.3956
	With Both (iPhone 13)	0.0271	0.0223	0.0174	0.0172	0.7461	0.6312

Comparing Gradient Boosting and SwinModel with and without preprocessing and augmentation, as shown in Table 14, it can be seen that modelling with preprocessing and/or augmentation performed better compared to the modelling without preprocessing and augmentation. For Gradient Boosting, modelling with preprocessing and augmentation using image captured using iPhone 13 performed the best. However, the R2 is better for modelling with preprocessing only or with augmentation only. For SwinModel, it is shown that modelling with preprocessing and augmentation performed significantly better compared to the modelling without the two processes. Modelling with preprocessing and augmentation using image captured using iPhone 13 even performed the best compared to the modelling of other methods, combination of processes and smartphone types.

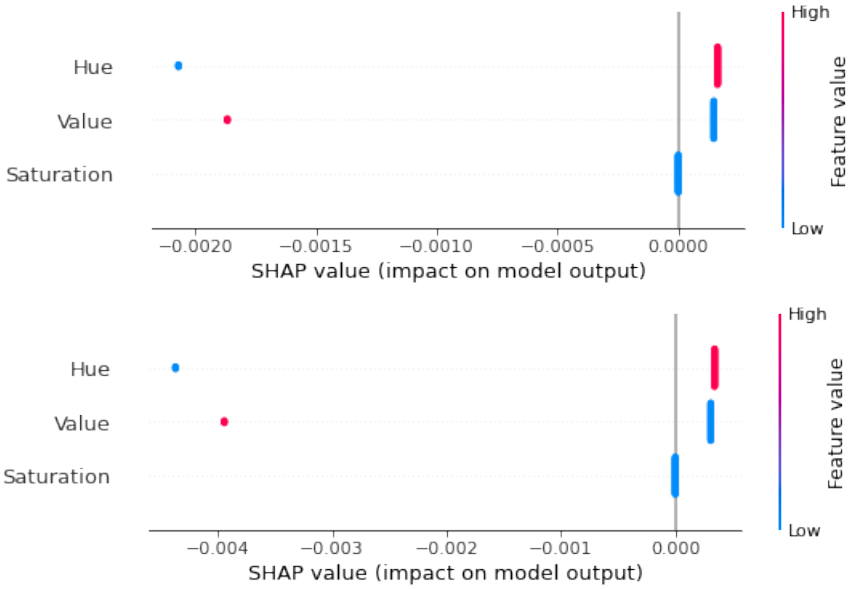

Figure 3. SHAP Values of Variables in Gradient Boosting Modelling for NPK (above) and pH (below)

To look at the effect of each variable in dataset toward the resulting model, Figure 3 shows the SHAP values of each variable in Gradient Boosting modelling. Hue shows the highest influence among the others and hence has

the strongest impact. Value has a high value which pushes the prediction higher, whereas hue has low value which decreases predictions. SHAP values for modelling NPK and pH show similar patterns with Saturations have the lowest impact in the two modellings.

Comparing the evaluation metric of squared errors between Gradient Boosting modelling and SwinModel, modelling, the t-statistic for the NPK modelling for the two methods is 1.8477 and p-value of 0.0775. For pH modelling, the t-statistic is 1.3522 and p-value is 0.1895. With these values, it can be concluded that the two modellings resulted in significant differences of squared errors. SwinModel has been remarkably better in performing soil image modelling and predicts the values of NPK and pH.

4. Discussion and Recommendation

4.1. Discussion

Modelling soil images to predict NPK and pH values requires a number of processes including preprocessing, augmentation, feature extraction, and classification modelling. For the modelling, various methods could be used including Gradient Boosting, Random Forest, ANN, and CNN. For ANN and CNN, a number of parameters need to be set for the training including batch size, epoch, neuron number, and optimizer. For CNN, an additional procedure of transfer learning could be implemented, to help CNN perform the training faster and obtain better results. Transfer learning could include ViTModel, SwinModel, ResNet50, and EfficientNetB4.

For the preprocessing, each method shows different tendencies. As shown in the experiments for soil NPK and pH modelling, some stood out preprocessing methods for soil analysis include median filter and homomorphic filter. Median filter is beneficial for soil analysis with its characteristics of noise removal and edge preservation [53]. Noise removal is needed for soil analysis, as soil images could consist of a lot of noises such as irrelevant objects, dust, and other noises. Edge preservation is also needed as the soil types, roots, and cracks still need to be retained. Median filters also have an advantage of ignoring extreme values, which can provide more reliable data. Homomorphic filters, on the other hand, have some advantages such as their capability to normalize brightness on soil images, moisture gradients enhancement, shadow effect reduction, and soil roughness enhancement, among others [54].

Other preprocessing methods such as white balance, replacing outliers, bilateral filter, and Gaussian blur, also showed reliable results for some setting of experiments [53]. These preprocessing methods also have their advantages for handling noises and brightness, which are beneficial for certain classification methods. At one experiment, modelling without preprocessing also produced satisfactory results when modelling NPK using ANN. These conditions could be due to the available data, as there are not too many shadows and irrelevant objects in the captured soil images, but they are still value variations on the dataset, as shown in the clustering results. ANN also known to be straight forward and can process raw pixel values directly without normalization or preprocessing, recognizing noises as flat feature without spatial invariance.

For the augmentation, color augmentation, physical augmentation, and rotation augmentation have shown good modelling results. Color augmentation is especially good for modelling NPK in ANN and CNN. Physical augmentation, on the other hand, performed well for all methods in modelling NPK and pH. Rotation augmentation, however, only suits Gradient Boosting and Random Forest methods.

Color augmentation is particularly good for soil analysis as it introduces variations of lighting, moisture, debris, and organic matter during the image preparation, making the dataset become invariant [55]. This also avoids modelling to experience overfitting, because of limitation of data. This is especially good for ANN and CNN, as it forces the feature to be spatially invariant, and makes the methods to learn global statistical patterns easily. Physical augmentation, on the other hand, replicates the existences of cracks, clumps, debris, and varying moisture levels [55]. Soil could also consist of diverse types of lands such as sandy, clayey, or loamy. This generalizes different soil conditions and improves modelling robustness. Rotation augmentation is particularly good for Gradient Boosting and Random Forest needs more data compared to ANN and CNN and the augmentation method can increase the availability of data needed. Unlike ANN or CNN, Gradient Boosting and Random Forest also use tabular feature inputs and need to diverse data feature manually.

For parameter setting in ANN and CNN modelling, as the number of available data is limited, the experiments show that lower batch sizes performed better. For the epoch and neurons, the highest number set are chosen also due to the availability of data in the data training, since for small dataset more epoch and neurons are needed to avoid overfitting. The agam optimizer is performed better than SGD (Stochastic Gradient Descent) since the agam optimizer is very flexible in setting the learning rate of the modelling. As soil data is noisy and complex, the type of optimizer that can be flexible following the conditions of data is needed [56]. Agam is also known to be better suited for small datasets because of its flexibility.

For transfer learning in CNN, ResNet50 performed the best compared to other transfer learning methods. This is due to its capabilities in handling small datasets [47]. As image data consists of noises and features, the method is suitable as it can recognize local patterns existing in the data. The other transfer learning methods, on the other hand, come second during the experiments as they have a number of specifications to be met such as the need of large amount of data, use complex scaling rules for EfficientNetB4 [48], need to implement advanced training tricks including warm-up and learning rate schedulers, for SwinModel and ViTModel [49, 50], and preference to be used for non natural science based structured data for SwinModel and ViTModel [49, 50].

Comparing the modelling of soil images captured using different smartphones based on selected set of processes shows that the patterns are the same with SwinModel performed the best and Gradient Boosting came second. However, the evaluation of metrics values is different between one to another as the specifications of cameras used to capture the images influence the resulting soil images used for modelling and prediction.

4.1.1. Recommendation Based on the results, soil images captured using smartphones could have large variations. This could be due to the conditions of soil which are affected by the crops that have been planted, the number of days in between farming processes, nutrient conditions, sunlight, temperature, and others. Based on the study, as many as eight variations were found in soil images. Treatment to soil nutrients based on soil color analysis could be conducted differently. Comparison between modelling images captured using several types of smartphones also shows that several types of smartphones result in different sets of soil images and consequently produce different accuracies in modelling and predicting the values of NPK and pH.

However, for the resulting classification model, different sets of soil imaged captured using several types of smartphones, do not alter the pattern of results produced from the modelling. Transformer types of deep learning performed better compared to conventional machine learning methods such as Gradient Boosting, Random Forest, and ANN. Variations of CNNs with inclusion of pretrained model also cannot compete with the modelling performed by Swin Transformer Model (SwinModel).

As for the preprocessing and augmentation needed before the modelling, it can be recommended to implement various preprocessing and augmentation methods before the process. For preprocessing, median filter and homomorphic filter are two preprocessing methods that could enhance the accuracy of modelling results. Augmentations based on color, physical, and rotation have also proven that they improved the quality of the modelling. As for the setting of modelling parameters such as batch size, number of epochs, image sizes, and optimizers, for batch size it will depends on the size of the soil dataset, number of epochs could be set to 100, image sizes are needed to be larger, and the most suitable optimizer is adam.

As the evaluation on aspects related to soil image modelling to predict NPK and pH values has been conducted, implementation in the form of mobile application is further recommended. The process of capturing images used in this study is set to be as natural as possible, so the farmer could easily imitate the process in the same way. The thing that the farmer needs to be aware of in implementing any mobile application for soil analyses is that the area of soil to be analyzed needs to be cleaned in a certain degree to have a bit clearer soil image for the analysis.

5. Conclusion

Several conclusions could be drawn from this study:

1. Soil images captured using smartphones could vary a lot. This is due to numerous factors such as soil nutrient, soil environment conditions, and smartphone types. The latter is even proven to influence the accuracy of modelling results.

2. Augmentation and preprocessing are still needed, even when we use deep learning methods. Color, physical, and rotation-based augmentation, and median and homomorphic filters are the methods that could be recommended for augmentation and preprocessing, respectively.
3. Transformer type of deep learning method such as SwinModel is good to be implemented for analysis of soil images in predicting the values of NPK and pH. However, conventional machine learning methods such as Gradient Boosting and Random Forest are also not far behind. With the speed of modelling, the use of the two latter methods is also recommended.
4. The study is suitable to be implemented into a mobile application used by farmers, as the soil images for the study were captured in a natural process which can be performed by farmers easily.

Acknowledgement

This work was supported by Institut Teknologi dan Bisnis STIKOM Bali (ITB STIKOM Bali) with the grant number of 530.136/DIRPPM&P/WRI/ ITBSTIKOM/WDS/XII/23.

REFERENCES

1. J. Harwood, R. Heifner, K. Coble, J, Perry, and A. Somwaru, *Managing Risk in Farming: Concepts, Research, and Analysis*, Agricultural Economic Report No. 774. , Market and Trade Economics Division and Resource Economic Division, Economic Research Service, U.S. Department of Agriculture, 1999.
2. B. Tedesse, Y. Tilahun, T. Bekele, and G. Mekonen, *Assessment of Challenges of Crop Production and Marketing in Bench-Sheco, Kaffa, Sheka, and West-Omo Zone of Southwest Ethiopia*, Heliyon 7, 2021.
3. M. I. Abdi-Soojeede, *Crop Production Challenges Faced by Farmers in Somalia: A Case Study of Afgoye District Farmers*, Wavelet Agricultural Science 9, vol 9, no 7, pp. 1032-1046, 2018.
4. V. K. H. Manegar, *Problems Faced by Farmers in Karnataka*, International Journal of Research and Analytical Reviews, vol 2, issue 4, pp. 46-56, 2015.
5. Z. Rozaki, *Food security challenges and opportunities in Indonesia post COVID-19*, Advances in Food Security and Sustainability, vol 6, pp. 199-168, 2021.
6. M. H. Shahapur, and R. T. Pote, *A Study on Problems Faced by Farmers of Yadgir District of Karnataka State*, International Journal of Creative Research and Thought, vol 9, issue 5, pp. 838-843, 2021.
7. D. Kahan, *Managing Risk in Farming*, Food and Agriculture Organization of The United Nations, Rome, 2008.
8. S. Sundaramoorthy, *A Study on Problems and Prospects of Farmers with Reference to Tirunelveli District*, International Journal of Economics, vol 9, issue 2, pp. 22-25, 2021.
9. UTASS and Rose Regeneration, *Challenges Facing Farmers*, A Report into Upland Farming and Farming Families in Teesdale, 2012.
10. J. F. Montanez, *Soil parameter detection of soil test kit-treated soil samples through image processing with crop and fertilizer recommendation*, Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, no. 1, pp. 90-98, 2021.
11. H. Pallevada, S. P. Potu, T. V. K. Munnangi, B. C. Rayapudi, S. R. Gadde, and M. Chinta, *Real-time Soil Nutrient detection and Analysis*, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, vol, pp. 1035-1038, 2021.
12. J. C. Puno, E. Sybongco, E. Dadios, I. Valenzuela, J. Cuello, *Determination of Soil Nutrients and pH level using Image Processing and Artificial Neural Network*, 9th IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1-6, 2017.
13. R. G. Regalado, J. C. D. Cruz, *Soil pH and Nutrient (Nitrogen, Phosphorus and Potassium) Analyzer using Colorimetry*, IEEE Region 10 Conference (TENCON), pp. 2387-2391, 2016.
14. R. Sudha, S. Aarti, K. Nanthini, *Determination of Soil Ph and Nutrient Using Image Processing*, International Journal of Computer Trends and Technology (IJCTT) – Special Issue, pp. 58-61, 2017.
15. B. H. Tan, W. H. You, S. H. Tian, T. F. Xiao, M. C. Wang, B. T. Zheng, L. N. Luo, *Soil Nitrogen Content Detection Based on Near-Infrared Spectroscopy*, Sensors, 22, 8013, 2022.
16. N. R. Zani, A. H. Alasiry, A. Wijayanto, *Design and Development of Soil Nutrients Level Detection System based on Soil Colour and pH for Crop Recommendations using Fuzzy Algorithms*, The Indonesian Green Technology Journal, pp. 38-45, 2022.
17. N. Moritsuka, K. Matsuoka, K. Katsura, J. Yanai. *Farm-scale variations in soil colour as influenced by organic matter and iron oxides in Japanese paddy fields*, Soil Science and Plant Nutrition, 65:2, pp. 166-175, 2019. DOI: 10.1080/00380768.2019.1583542.
18. M. Amri, R. Sumiharto, *Sistem Pengukuran Nitrogen, Fosfor, Kalium Dengan Local Binary Pattern Dan Analisis Regresi*, Indonesian Journal of Electronics and Instrumentation Systems (IJEIS), Vol. 9, No. 2, pp. 107¯118, 2019.
19. M. J. Aitkenhead, M. Coull, W. Towers, G. Hudson, H. I. J. Black, *rediction of soil characteristics and colour using data from the National Soils Inventory of Scotland*, Geoderma 200–201, pp. 99–107, 2013.
20. M. Mouazen, R. Karoui, J. Deckers, J. De Baerdemaeker, H. Ramon, *Potential of visible and near-infrared spectroscopy to derive colour groups utilising the Munsell soil colour chart*, Biosystems Engineering, vol 97, Issue 2, pp. 131-143, 2007.
21. R. A. V. Rossel, B. Minasny, P. Roudier, A. B. McBratney, *Colour space models for soil science*, Geoderma 133, pp. 320-337, 2006.

22. S. S. Nodi, M. Paul, N. Robinson, L. Wang, S. U. Rehman, *Determination of Munsell Soil Colour Using Smartphones*, Sensors, 23, 3181, 2023.
23. Y. Agusta and D. L. Dowe, *MML Clustering of Continuous-Valued Data Using Gaussian and t Distributions*, In Lecture Notes in Artificial Intelligence, vol. 2557, pp. 143-154, 2002.
24. Y. Agusta and D. L. Dowe, *Unsupervised Learning of Gamma Mixture Models Using Minimum Message Length*, 3rd IASTED International Conference on Artificial Intelligence and Applications, pp.457-462, 2003.
25. Y. Agusta and D. L. Dowe, *Unsupervised Learning of Correlated Multivariate Gaussian Mixture Models Using MML*, Lecture Notes in Artificial Intelligence AI2003, vol. 2903, pp. 477-489, 2003.
26. Y. Agusta, *Implementing Minimum Message Length to the Modelling of Denpasar City Inflation Rate*, 2023 Eighth International Conference on Informatics and Computing (ICIC), pp. 1-6, 2023.
27. C. S. Wallace and D. L. Dowe, *Intrinsic classification by MML – the Snob program*, In Proc. 7th Aust. Joint Conf. on AI, pp. 37-44, 1994.
28. H. Robbani, E. T, Trisnawati, R. Noviyanti, A. Rivaldi, F. P. Cahyani, F. Utaminingrum, *Aplikasi Mobile Scotect: Aplikasi Deteksi Warna Tanah Dengan Teknologi Citra Digital Pada Android*, Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), Vol. 3, No. 1, pp. 19-26, 2016.
29. P. Raj, A. Kumar, R. Singh, V. Patel, *Adaptive filtering for soil image noise removal*, Sensors, 21(9), 3086, 2021. DOI: 10.3390/s21093086.
30. Y. Liu, H. Wang, X. Chen, K. Zhang, *Pre-segmentation noise reduction in soil images using Gaussian filtering*, Sensors, 21(14), 4728, 2021. DOI: 10.3390/s21144728.
31. S. Lee, M. Kim, J. Park, H. Choi, *Learnable bilateral filters for soil image enhancement*, IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15, 2024. DOI:10.1109/TGRS.2023.3321541.
32. A. Kumar, B. Silva, *Benchmarking White Balance Methods for Precision Agriculture*, Journal: Biosystems Engineering, 208, 1-15, 2021. DOI: 10.1016/j.biosystemseng.2021.05.003.
33. L. Zhang, Y. Li, Y. Zhang, J. Wang, *Shadow Removal in Field Soil Images Using MSRCR*, Computers and Electronics in Agriculture, 187, 106301, 2021. DOI:10.1016/j.compag.2021.106301
34. L. Zhang, Y. Li, Y. Zhang, J. Wang, *Shadow removal in field soil images using optimized homomorphic filtering*, Biosystems Engineering 208, 1-14, 2021. DOI:10.1016/j.biosystemseng.2021.04.003
35. C. Shorten, T. M. Khoshgoftaar, *A survey on Image Data Augmentation for Deep Learning*, J Big Data 6, 60, 2019.
36. Y. Chen, L. Wang, J. Zhang, X. Li, *Color augmentation for small soil datasets*, Computers and Electronics in Agriculture 178, 105731, 2020. DOI:10.1016/j.compag.2020.105731.
37. L. Zhang, Y. Li, J. Wang, H. Chen, *Physics-based soil moisture augmentation*, Water Resources Research 57(8), e2020WR029456, 2021. DOI:10.1029/2020WR029456.
38. R. Kumar, A. Sharma, V. Patel, S. Singh, *Learned rotation augmentation for soil analysis*, IEEE Geoscience and Remote Sensing Letters 19, 1-5, 2022. DOI:10.1109/LGRS.2021.3136785
39. L. Zhang, J. Wang, Y. Li, H. Chen, *Physics-based wet soil appearance simulation*, IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 6, pp. 5181-5192, 2021. DOI: 10.1109/TGRS.2020.3026051.
40. L. Wang, Y. Zhang, X. Chen, J. Li, *ANN Architecture for Soil Organic Carbon Prediction*, Computers and Electronics in Agriculture 180, 105890, 2020. DOI: 10.1016/j.compag.2020.105890.
41. J. Smith, A. Johnson, B. Williams, R. Davis, *Optimization of batch sizes for soil property prediction using deep learning*, Computers and Electronics in Agriculture, 193, 106712, 2022. DOI: 10.1016/j.compag.2022.106712.
42. M. Garcia, P. Rodriguez, J. A. Martinez-Lopez, R. Fernandez, *Early Stopping Criteria for Soil Moisture ANNs*, Geoderma 405, 115400, 2022. DOI: 10.1016/j.geoderma.2021.115400.
43. P. Zhou, X. Li, Y. Wang, H. Zhang, *Optimizer Benchmarking in Remote Sensing*, Remote Sensing 15(3), 712. 2023. DOI: 10.3390/rs15030712.
44. L. Kalyani, K. B. Prakash, *Soil Color as a Measurement for Estimation of Fertility using Deep Learning Techniques*, International Journal of Advanced Computer Science and Applications, vol. 13, no. 5, pp. 305-310, 2022.
45. F. Chollet, *Deep Learning with Python*, New York: Manning Publication. 2018.
46. Y. Goodfellow, A. Bengio, Courville, *Deep Learning*, MIT Press. 2016.
47. K. He, X. Zhang, S. Ren, J. Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. 2016.
48. L. Zhang, J. Wang, Y. Li, H. Chen, *EfficientNetB4 for Remote Sensing: A Benchmark*, IEEE ransactions on Geoscience and Remote Sensing, 59(8), 6821-6833, 2021.
49. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 9th International Conference on Learning Representations (ICLR 2021), 2021. https://openreview.net/forum?id=YicbFdNTTy.
50. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), pp. 10012-10022, 2021. DOI: 10.1109/ICCV48922.2021.00986.
51. S. Dharumarajan and R. Hedge, *Digital mapping of soil texture classes using Random Forest classification algorithm*, Soil Use and Management. British Society of Soil Science, Wiley, 2020.
52. Q. T. Bui, T. Y. Chou, T. V. Hoang, Y. M. Fang, C. Y. Mu, P. H. Huang, V. D. Pham, Q. H. Nguyen, D. T. N. Anh, V. M. Pham, M. E. Meadows, *Gradient Boosting Machine and Object-Based CNN for Land Cover Classification*, Remote Sens, 13 (14), 2709, 2021.
53. B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, A. Sharma, *Image denoising review: From classical to state-of-the-art approaches*, Information Fusion, Volume 55, March 2020, Pages 220-244. DOI: 10.1016/j.inffus.2019.09.003.
54. M. Oktiana, T. Horiuchi, K. Hirai, K. Saddami, F. Arnia, Y. Away, K. Munadi, *Cross-spectral iris recognition using phase-based matching and homomorphic filtering*, Heliyon. 2020 Feb 20; 6 (2): e03407. DOI: 10.1016/j.heliyon.2020.e03407.

55. L. Nanni, M. Paci, S. Brahnam, A. Lumini, *Comparison of Different Image Data Augmentation Approaches*, J Imaging. 2021 Nov 27; 7(12): 254. DOI: 10.3390/jimaging7120254.
56. J. Wang, J. Wiens, *AdaSGD: Bridging the Gap Between SGD and Adam*, International Conference on Machine Learning (ICML), 2020. DOI: 10.48550/arXiv.2006.16541