MIS-ME: A Multi-modal Framework for Soil Moisture Estimation*

Mohammed Rakib*, Adil Aman Mohammed*, Cole Diggins†, Sumit Sharma†,
Jeff Michael Sadler†, Tyson Ochsner†, Arun Bagavathi*
*Department of Computer Science, †Department of Plant and Soil Sciences
Oklahoma State University, Stillwater, Oklahoma, United States
{mohammed.rakib, adil.mohammed, ddiggin, sumit.sharma, jeff.sadler, tyson.ochsner, abagava}@okstate.edu

***Abstract*—Soil moisture estimation is an important task to enable precision agriculture in creating optimal plans for irrigation, fertilization, and harvest. It is common to utilize statistical and machine learning models to estimate soil moisture from traditional data sources such as weather forecasts, soil properties, and crop properties. However, there is a growing interest in utilizing aerial and geospatial imagery to estimate soil moisture. Although these images capture high-resolution crop details, they are expensive to curate and challenging to interpret. *Imagine*, an AI-enhanced software tool that predicts soil moisture using visual cues captured by smartphones and statistical data given by weather forecasts. This work is a first step towards that goal of developing a multi-modal approach for soil moisture estimation. In particular, we curate a dataset consisting of real-world images taken from ground stations and their corresponding weather data. We also propose *MIS-ME* - Meteorological & Image based Soil Moisture Estimator, a multi-modal framework for soil moisture estimation. Our extensive analysis shows that MIS-ME achieves a MAPE of 10.79%, outperforming traditional unimodal approaches with a reduction of 2.6% in MAPE for meteorological data and 1.5% in MAPE for image data, highlighting the effectiveness of tailored multi-modal approaches.**

***Index Terms*—multi-modal regression, visual feature extraction, soil-moisture estimation**

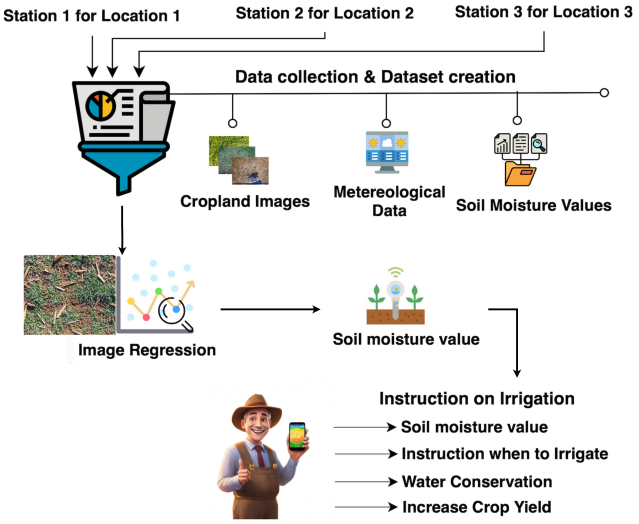

Fig. 1: Integrating cropland images and meteorological data for improved soil moisture estimation and irrigation guidance.

I. Introduction

Soil moisture is a key indicator of water usage and availability in agricultural fields [1]. Soil moisture levels determine the amount of irrigation that is needed in irrigated fields, and it informs yield prediction in rain-fed fields. When there is insufficient soil moisture, crop productivity suffers, causing a significant financial burden for farmers. Although crucial, accurately measuring soil moisture requires in-situ sensor installations, which are difficult to scale and introduce upfront and ongoing maintenance costs [2]–[4]. These difficulties are accentuated by ongoing efforts to develop and adopt precision agriculture. Precision agriculture, which uses technology to assist with sustainable agriculture management, though not a new idea [5], has gained increased traction in the past decade with advances in data availability, hardware (internet-connected devices), and data algorithms [6]. AI-based data-driven methods are one of the precision agriculture techniques that has proven useful in multiple applications like soil management, optimal crop-producing conditions, and determining watering quantities and fertilizers [7]. Developing such data-driven models for accurate soil moisture predictions, therefore, could provide useful information to farmers regarding locations that are unmonitored. In this work, we focus on utilizing machine learning approaches with digital images and weather observations collected from ground stations to estimate soil moisture to advance AI-driven precision agriculture.

Soil moisture prediction has become a ubiquitous sub-task in the computational agriculture domain with the advancements in machine learning and deep learning algorithms [8], [9]. Multiple applied research directions for utilizing machine learning approaches in the soil moisture prediction sub-task have been explored in the literature. The most common approach is to utilize meteorological data, weather forecasts, geographical details, and soil properties to predict the soil moisture level ahead in the future [8], [10], [11]. These methods utilize traditional machine learning models like linear regression, neural network regressor, and random forest models for such predictions. The second spectrum of research focuses on soil moisture prediction on either satellite imagery [12] or UAV-based multispectral and thermal images [13]. Such images offer a high-level understanding of the farmland and

*This work is funded by USDA: Agriculture and Food Research Initiative (AFRI) - Data Science for Food and Agriculture Systems (DSFAS); Award Number:2023-67022-40019

capture macro details of the area. However, these datasets are costly to procure, often capture only information about soil, and do not characterize micro details about crops and their relation to the soil moisture. Some of the works that study such micro details in soil moisture prediction sub-tasks consider images collected from a lab setting, which are less noisy [14]–[16]. Estimating soil moisture using photos taken from mobile phones or field cameras could be an impactful development towards real-time precision agriculture. As depicted in Fig. 1, we present an initial work towards precision agriculture with multi-modal data and develop methods to understand patterns in soil moisture to assist farmers in real time. Particularly, we contribute to analyzing the capacities of machine learning models on soil moisture estimation tasks directly on images taken from field cameras along with meteorological weather observations. We study soil moisture in terms of volumetric water content (*VWC*), which is the proportion of water volume to soil volume, usually expressed as cm^3cm^{-3} [17].

We present ***three-fold*** contributions in this work:

- We curate soil patch images from the existing photos collected from ground stations along with associated meteorological data, with an aim to enhance the performance of soil moisture or *VWC* estimation problems.
- We evaluate the potential of image regression models to predict the *VWC* from our pre-processed datasets. To our knowledge, this is the first effort to quantify soil moisture directly from raw cropland images.
- We introduce the *MIS-ME* framework, featuring three innovative multi-modal approaches for soil moisture estimation: MIS-ME with Multi-modal Concat, MIS-ME with Hybrid Loss, and MIS-ME with Learnable Parameters. Each approach is designed to leverage both image features from soil patches and corresponding meteorological data, optimizing prediction accuracy through distinct combination and weighting strategies. The code will be made available after the review process.

II. Related Work

A. Meteorological Data in Soil Moisture Prediction

Current trends in soil moisture forecasting display a broad spectrum of applications for meteorological data, ranging from statistical models to deep learning methodologies. The Seasonal ARIMA model, when coupled with a water balance equation, shows significant accuracy in predicting soil moisture across various depths [18], proving effective especially at depths minimally affected by external variables. Moving to machine learning, [19] employs multiple linear regression and support vector regression, showcasing the effectiveness of linear regression in specific contexts. LightGBM's potential in high-resolution soil moisture prediction outperforms traditional methods like linear regression and random forest, especially in IoT-enabled soil moisture data scenarios [20].

Deep learning has been widely used for soil moisture prediction with models like LSTMs, 1D-CNNs, encoders, and even the fusion of CNNs with LSTMS [21]. These models effectively handle diverse soil moisture data under varying conditions. [22] proposes a Deep Neural Network Regression (DNNR) model, noted for its robust data fitting and generalization capabilities in soil moisture prediction. Additionally, [23] introduces the EDT-LSTM model, enhancing prediction accuracy by focusing on intermediate time-series data. Similarly, [10] presents a hybrid CNN-GRU model tailored for maize root zone moisture prediction, which excels in accuracy and convergence rate. Collectively, these studies underscore the rapidly evolving landscape of soil moisture prediction methodologies, each contributing unique insights and advancements to the field.

B. Image-Based Soil Moisture Estimation

The application of satellite and UAV imagery in soil moisture prediction has revolutionized the field, utilizing various methodologies to improve precision and utility. The first approach involves extracting features from satellite images for soil moisture prediction. Studies like [24] and [12] demonstrate this technique by integrating features from Sentinel-1 backscatter data, soil moisture active passive data, and topographic information within deep learning frameworks. Similarly, [25] focuses on the effective use of Sentinel-2 bands and indices like NDWI and NDVI, utilizing a convolutional neural network for accurate soil moisture estimation. Additionally, [26] introduces models based on ConvLSTM layers and visual transformers, employing NDVI and NSMI satellite data for soil moisture content prediction, highlighting the effectiveness of deep learning in processing satellite-derived information. In the second approach, UAV imagery is employed for soil moisture prediction, offering high-resolution data essential for precision agriculture [9], [27]. [9] explores the use of visible UAV imagery combined with models like Random Forest and Multilayer Perceptron, demonstrating the potential of aerial imagery in bare soil field assessments. Lastly, lab-based image analysis offers a controlled environment for soil moisture prediction. [14], [15], and [16] utilize soil surface images within laboratory settings to develop predictive models. These studies apply image processing techniques and convolutional neural networks to analyze soil water content and density, showcasing the capabilities of image processing in soil analysis under controlled conditions.

C. Our Contribution

In contrast to existing research primarily relying on controlled lab settings for soil image acquisition, our analysis uses raw images captured in natural environments such as field node cameras. Our methodology stands out by integrating these real-world soil images with corresponding meteorological data to enhance the prediction of soil moisture. Moreover, we introduce the three-way *MIS-ME* framework, which presents three fusion techniques to combine soil patch images and tabular meteorological data. Our method of utilizing raw soil images captured in the wild positions our research as a distinctive and impactful contribution towards precision agriculture.

TABLE I: Statistical overview of the soil moisture dataset

Station Name	Sand-Silt-Clay (%)	Crop land Images	Soil Patches	Days Count	*VWC* Range (cm^3cm^{-3})	*VWC* Mean (cm^3cm^{-3})	*VWC* STD (cm^3cm^{-3})
Station1	18.0 - 56.5 - 25.6	977	1822	202	0.158 - 0.417	0.3085	0.0496
Station2	28.8 - 45.2 - 26.0	704	1604	246	0.118 - 0.435	0.2455	0.0646
Station3	19.5 - 54.2 - 26.3	925	3356	258	0.151 - 0.423	0.3126	0.0582

TABLE II: Evaluation metrics for YOLOv5 model

Data Split	Evaluation criteria	Precision	Recall	F1 Score	AP
Train	IoU@0.5	0.830	0.706	0.763	0.756
	IoU@0.75	0.830	0.706	0.763	0.754
	IoU@0.90	0.801	0.688	0.740	0.728
Validation	IoU@0.5	0.830	0.706	0.763	0.756
	IoU@0.75	0.830	0.706	0.763	0.754
	IoU@0.90	0.801	0.688	0.740	0.728
Test	IoU@0.5	0.818	0.685	0.745	0.758
	IoU@0.75	0.805	0.682	0.738	0.756
	IoU@0.90	0.739	0.659	0.697	0.710

III. Dataset

In this work, we use field camera images and associated meteorological data for machine-learning tasks. The dataset was collected as a part of ongoing research to develop improved cropland monitoring stations. These stations are equipped with cosmic-ray neutron detectors for non-contact, field-scale soil moisture monitoring, downward-facing outdoor cameras for visually monitoring soil and crop conditions, and all-in-one weather stations for recording meteorological data. These stations were deployed in three agricultural fields in the Southern Great Plains during 2020-2021 to collect atmospheric data and cropland images. The meteorological data include air temperature, relative humidity, and rainfall. Images were collected at regular intervals from 9 am to 5 pm. These data capture the dynamic nature of soil moisture, reflecting the complexity and variability of agricultural croplands.

Table I presents a statistical overview of the raw dataset, encompassing data from three stations used in this work. The table specifies the respective soil compositions in terms of sand, silt, and clay. Notably, all stations exhibit similar clay content, with marginal variations in sand and silt ratios, particularly at *Station2*. Additionally, the table illustrates the range, mean, and standard deviation of volumetric water content (*VWC*) at each station. The *VWC* across the stations varies from approximately 0.1 to 0.4 cm^3cm^{-3}. Both *Station1* and *Station3* show similar mean values and standard deviations for soil moisture, whereas *Station2* exhibits a marginally lower mean and the highest standard deviation. This indicates that while *Station1* and *Station3* share comparable soil characteristics, *Station2* differs slightly in its soil properties.

A. Preprocessing Cropland Images for Soil Patch Extraction

Images available in our dataset are collected from crop fields, unlike existing images-based soil moisture prediction methods [14]–[16], which collect soil samples from a lab setting. Hence, our raw cropland images include not only soil patches but also objects like crops, shadows, blurred images, and night images. Initially, we had 1950 images for *Station1*, 1404 images for *Station2*, and 1835 images for *Station3*. After manually removing noisy images, we have a total of 2606 raw images with 977 images from *Station1*, 704 images from *Station2*, and 925 images from *Station3*. In this work, we follow a 3-step approach to prepare these filtered cropland images for machine learning tasks: (i) handpick and label (draw bounding box over soil patches) 250 cropland images, (ii) train & evaluate YOLOv5 [28] using those images and (iii) extract soil patches using the trained YOLOv5 from the filtered images of the three stations. Fig. 2 gives a complete overview of the dataset creation process.

First, we manually pick *260* raw images and draw bounding boxes to mark soil patches. This small sample is split into three categories: train (210 images), validation (25 images), and test (25 Images) sets. Next, we train and evaluate the YOLOv5 object detection model to extract soil patches from all preprocessed images. Table II shows the performance of the finetuned YOLOv5 model with Intersection over Union (IoU) evaluation criteria thresholds at 0.5, 0.75, and 0.90. Observing Table II, we see that the model performs well across all overlapping criteria for the training, validation, and test sets. The precision, recall, F1 score, and AP remain steady and show similar trends across all the dataset splits. The evaluation results indicate that the trained YOLOv5 model is effective in detecting soil patches with consistent performance across different IoU thresholds and data splits. Subsequently, we use this trained *YOLOv5* model to extract a total of 6782 soil patches from all our preprocessed 2606 raw images across the three stations and create the dataset. We set a confidence score of 0.5 for the YOLOv5 model to ensure that the model is at least 50% sure when detecting the soil patches and to obtain a good quantity of data for machine learning models. After extracting soil patches and eliminating irrelevant images,

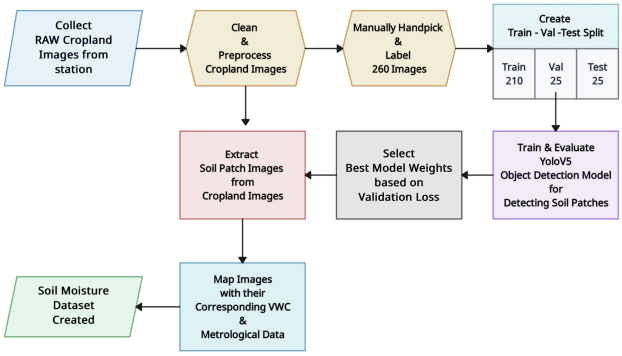

Fig. 2: Overview of the dataset creation pipeline

TABLE III: Description of Meteorological Variables

Abbreviation	Variable Description	Unit
T_{air}	air temperature	°C
T_{mod}	temperature of the internal system module	°C
T_{hs}	temperature of the humidity sensor	°C
RH	percent relative humidity	%
RH_{mod}	humidity of the internal system module	%
P	precipitation	mm
Φ_{solar}	solar radiation flux given by ClimaVUE 50 sensor	W/m^2
P_{vapor}	pressure exerted by water vapor in air	kPa
P_{bar}	barometric pressure	hPa
v_{wind}	wind speed	m/s
v_{gust}	highest speed of wind during short bursts	m/s
v_{north}	north relative wind speed	m/s
v_{east}	east relative wind speed	m/s
θ_{wind}	direction of the wind	degrees
$Tilt_{\text{NS}}$	north south tilt of the station	degrees
$Tilt_{\text{WE}}$	east west tilt of the station	degrees

Station3 has the highest soil patch quantity with 3356 patches, followed by 1822 patches for *Station1* and 1604 for *Station2*. ***We will make these pre-processed soil patches publicly available after the peer review process.***

B. Preprocessing Meteorological Data

Table III provides an overview of the meteorological variables measured by the cropland monitoring stations with their respective units. While each variable offers valuable insights for meteorological studies, our research will specifically focus on variables that strongly correlate with *VWC*. Fig. 3 displays the Pearson correlation analysis between *VWC* and various meteorological variables. Notably, *VWC* exhibits significant inverse correlations with T_{air}, T_{mod}, and T_{hs}, with coefficients of -0.28, -0.26, and -0.27, respectively. In contrast, *VWC* has a strong positive correlation of 0.38 with RH. Due to the high intercorrelation among T_{air}, T_{mod}, and T_{hs}, we retain only T_{air} and exclude the others in our dataset to enhance machine learning efficiency and reduce redundancy. *VWC* also correlates linearly with P_{bar}, Φ_{solar}, $Tilt_{NS}$, $Tilt_{WE}$, P, v_{wind}, and v_{gust}, with respective coefficients of 0.18, 0.14, 0.09, 0.13, 0.08, 0.08, and 0.09. Due to the high correlation between v_{wind} and v_{gust}, we choose to include only v_{wind}. We exclude other variables, such as RH_{mod}, P_{vapor}, v_{north}, v_{east}, and θ_{wind} due to minimal correlation with *VWC*.

The selected meteorological features for our model training are T_{air}, RH, P, P_{bar}, Φ_{solar}, $Tilt_{NS}$, $Tilt_{WE}$, and v_{wind}, focusing on the variables that most significantly impact *VWC*. All the selected features are normalized using the standard z-score normalization technique.

C. Dataset Analysis

In our analysis of the *VWC* in Fig. 4, we note that *Station1* and *Station3*, with similar soil types, show aligned *VWC* peaks, likely linked to similar reactions to moisture changes during events like irrigation or rainfall. On the other hand, *Station2*, characterized by a higher sand content, displayed sharper *VWC* peaks and rapid declines, highlighting the influence of its soil texture on moisture fluctuations. This pattern, especially evident in *Station2*, emphasizes the role of soil texture in managing water retention and drainage.

IV. Methodology

In this section, we formulate and describe the traditional unimodal approaches and our proposed three-way MIS-ME framework for soil moisture estimation.

A. Image-Only Regression Approach

This approach utilizes state-of-the-art image feature extractors such as ResNet18 [29], InceptionV3 [30], MobileNetV2 [31], and EfficientNetV2 [32], each offering unique advantages for image regression tasks. The image feature extraction process can be represented by Eq. 1.

$$ImageFeatureExtractor(X_{img}) : \mathcal{F}(X_{img}) \rightarrow \mathbb{R}^n \quad (1)$$

Here, $X_{img} \in \mathbb{R}^{w \times h \times 3}$ denotes the input soil patch image, $\mathcal{F}$ represents the feature extraction function and it can be any of the state-of-the-art feature extraction models. $\mathbb{R}^n$ is an $n-$dimensional feature vector of X_{img}. The extracted features are then mapped to the predicted *VWC* value through a dense output layer, as given in Eq. 2.

$$ImageRegressionLayer(\mathbb{R}^n) : f(\mathbb{R}^n) \rightarrow \mathbb{R} \quad (2)$$

where, f denotes the linear regression function.

B. Meteo-Only Regression Approach

This approach focuses on tabular meteorological data only for *VWC* prediction. We introduce our own architecture for this approach titled Meteorological Soil Moisture Estimator (MSME). MSME has a series of fully connected layers with dropout and batch normalization to process the meteorological data efficiently. It is used in the MIS-ME framework to train

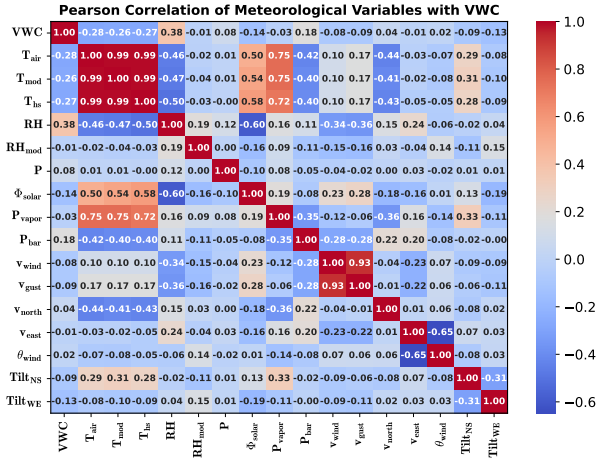

Fig. 3: Pearson correlation heatmap of the 16 meteorological variables with *VWC*.

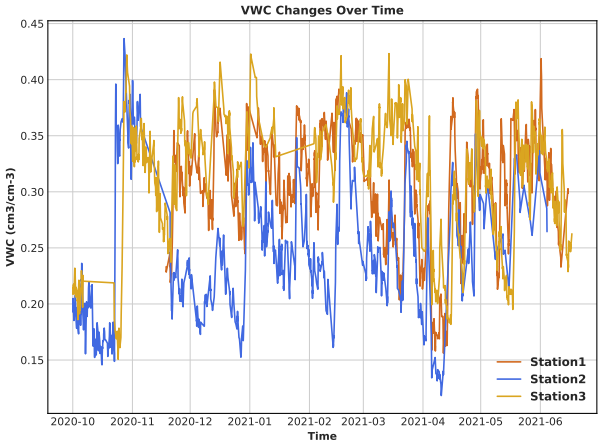

Fig. 4: Trend of *VWC* from October 2020 to June 2021 of the three stations with *Station1* and *Station3* showing similar trends compared to *Station3*.

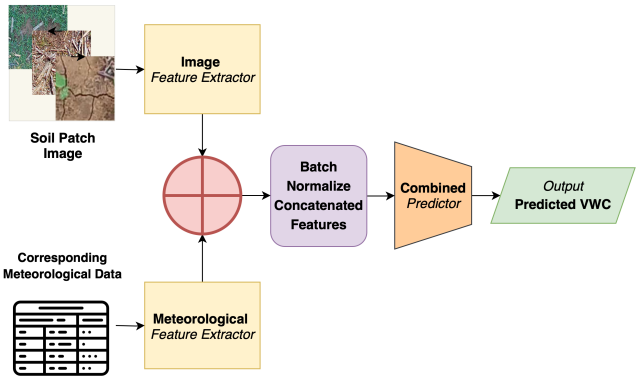

Fig. 5: MIS-ME with Multimodal Concat extracts soil patch image features using a trained image feature extractor and extracts tabular meteorological features using the proposed MSME model in IV-B. The extracted features are combined and batch-normalized to pass through a series of fully connected layers for soil moisture regression.

on meteorological data. The meteorological feature extraction process for MSME is given in Eq. 3.

$$MSME(X_{met}) : \mathcal{M}(X_{met}) \rightarrow \mathbb{R}^m \tag{3}$$

Here, $X_{met} \in \mathbb{R}^k$ is the input meteorological data vector, $\mathcal{M}$ represents sequential neural network operations, and m is the number of extracted features. These features are then fed into a linear regression layer to predict *VWC*, as given in Eq. 4.

$$MeteoRegressionLayer(\mathbb{R}^m) : g(\mathbb{R}^m) \rightarrow \mathbb{R} \tag{4}$$

where, g is the linear regression function.

C. Meteorological & Image based Soil-Moisture Estimator (MIS-ME) Framework

We propose the MIS-ME framework to combine both image and meteorological data for enhanced knowledge-encoded *VWC* prediction. We introduce three novel multimodal approaches for our MIS-ME framework, that have not been explored before in the field of soil moisture estimation. The approaches are formalized below utilizing formulations given in Eq. 1 and Eq. 3.

1) MIS-ME with Multimodal Concat: The concatenation approach within the MIS-ME framework directly combines the feature vectors obtained from both soil patch images and meteorological data. This method leverages the diverse feature representations directly for regression. The concatenated vector is then normalized and fed into a sequence of fully connected layers to predict the *VWC*. The mathematical representation of the concatenation process is shown in Eq. 5.

$$\begin{aligned} &MISME_{concat}(X_{img}, X_{met}) : \\ &\mathcal{P}(\mathcal{F}(X_{img}) \oplus \mathcal{M}(X_{met})) \rightarrow \mathbb{R}^{m+n} \end{aligned} \tag{5}$$

Here, $X_{img} \in \mathbb{R}^{w \times h \times 3}$ and $X_{met} \in \mathbb{R}^k$ represent the input image and meteorological data respectively. The feature vectors $\mathcal{F}(X_{img})$ and $\mathcal{M}(X_{met})$ are extracted using their respective feature extractors. The $\oplus$ operator denotes the concatenation of these vectors, combining them into a single feature vector. Next, we apply batch normalization to align the scale of the features from the two different sources. This is done to stabilize and smoothen the learning process of the model. The normalized combined features are then passed to the prediction function $\mathcal{P}$, which applies a series of dense layers to output the predicted *VWC*.

Fig. 5 illustrates the data flow through the model, showing how inputs are transformed into a single output through the concatenation and subsequent processing steps. This setup allows the model to learn from both types of data simultaneously, potentially capturing interactions between image-derived features and meteorological factors that are indicative of soil moisture.

2) MIS-ME with Hybrid Loss: This approach employs a unique strategy by integrating three different loss functions designed to enhance the model's predictive accuracy and robustness across different types of input data. This multi-loss strategy harmonizes the losses from concatenated features, meteorological features, and image features, each weighted to optimize their contribution towards the final prediction. As illustrated in Fig. 6, this approach facilitates a holistic learning process that captures diverse characteristics of both soil patch images and meteorological data. The mathematical formulation of our triplet loss is given in Eq. 6.

$$\begin{aligned} \mathcal{L}_{hybrid} = \delta \cdot \mathcal{L}_{concat}(cO, GT) + \gamma \cdot \mathcal{L}_{meteo}(mO, GT) + \\ (1 - \delta - \gamma) \cdot \mathcal{L}_{image}(iO, GT) \end{aligned} \tag{6}$$

where:

- $\mathcal{L}_{concat}$ is the loss computed from the concatenated output (cO) against the ground truth (GT).
- $\mathcal{L}_{meteo}$ is the loss from the meteorological output (mO).
- $\mathcal{L}_{image}$ is the loss calculated from the image output(iO).

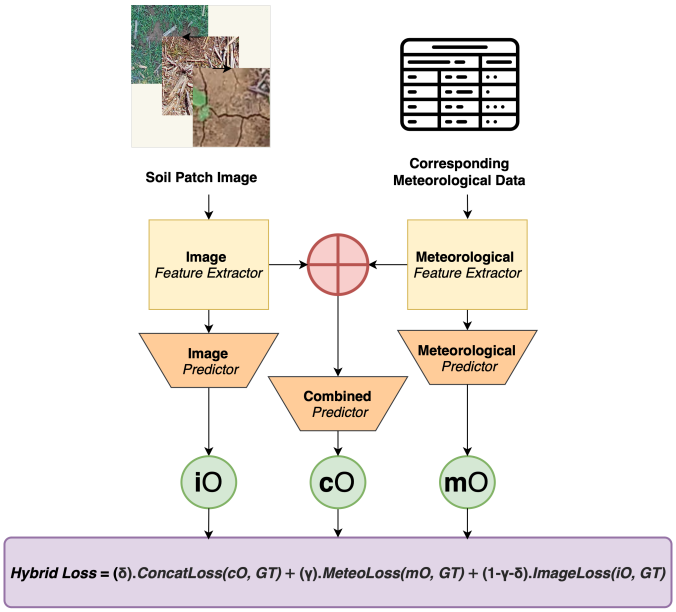

Fig. 6: MIS-ME with Hybrid Loss differs from the Multimodal Concat approach by introducing two more predictors apart from the combined predictor and a hybrid loss function that combines multiple losses in a weighted manner.

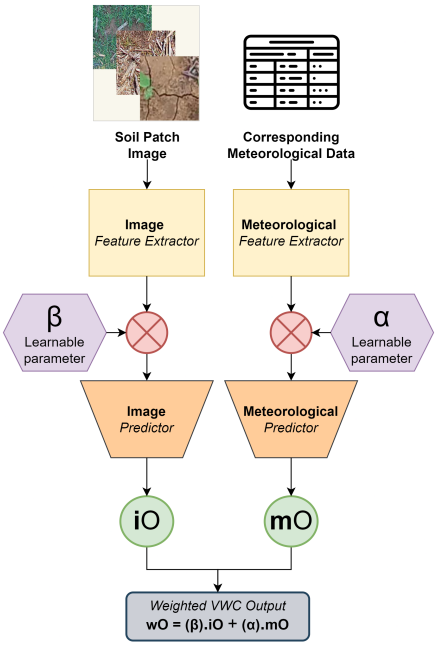

Fig. 7: MIS-ME with Learnable Parameter introduces two learnable parameters: β for images and α for meteo data, each of which gets multiplied with the output from their respective feature extractors. The final *VWC* output is then weighted using the learned value of these two parameters.

- δ, γ, and $1-\delta-\gamma$ are the weighting coefficients for the concatenated loss, meteorological loss, and image loss, respectively.

To the best of our knowledge, this application of a triplet loss framework, incorporating a dynamic weighting of losses across multiple modalities, represents a novel approach in the field of multimodal soil moisture estimation. By leveraging this method, our model not only adapts to the inherent variability in the data sources but also significantly enhances the robustness of the predictions by systematically balancing the influence of different feature sets.

3) MIS-ME with Learnable Parameter: This approach utilizes two distinct learnable coefficients, α and β, to dynamically adjust the influence of meteorological and image-derived features on the final prediction, as shown in Fig. 7. This method enables the model to adaptively allocate more weight to the modality that provides more predictive power, optimizing the integration of different data sources for soil moisture estimation. The formulation for the model prediction using two learnable parameters is expressed in Eq. 7.

$$\hat{y} = \alpha \cdot \mathcal{P}_{meteo}(mO) + \beta \cdot \mathcal{P}_{image}(iO) \tag{7}$$

where:

- $\mathcal{P}_{meteo}(mO)$ represents the prediction derived using meteorological data
- $\mathcal{P}_{image}(iO)$ denotes the prediction derived using image data
- α and β are the learnable parameters that get multiplied with the output of the respective feature extractors before passing through the predictors in the forward pass. This is done to ensure that the parameters become learnable by getting their gradients updated. In the end, the updated α and β parameters are multiplied with the predictions from the image and meteorological predictors, respectively, to get the final weighted *VWC* as output.

This method is designed to allow the model to automatically focus more on the data modality that is performing better, thereby adapting the prediction mechanism based on the intrinsic value of each type of data. By optimizing α and β within the training process, the model flexibly adapts to the underlying patterns and relationships in the dataset, enhancing both accuracy and robustness in predictions.

To the best of our knowledge, this approach of using dual learnable parameters for dynamic feature weighting has not been explored in previous studies. This approach not only boosts the model's adaptability and performance but also provides deeper insights into the relative importance of different data sources in predicting soil moisture. Fig. 7 illustrates the adaptive weighting mechanism, highlighting how α and β influence the final prediction based on the effectiveness of the data modalities.

V. Results & Discussion

In this section, we evaluate our proposed three-fold MIS-ME framework with existing baselines leveraging evaluation metrics like Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

A. Dataset Splitting Scheme

We have combined the data from all three stations to form the training, validation, and test sets in the ratio 65:15:20. Apart from evaluating the combined test set, the 20% test samples from each station have been evaluated separately as well to get an insight into the station-wise performance of the model. The validation set has played a crucial role in the hyperparameter tuning process of the models.

B. Performance Analysis of Meteo-Only Models

We assess the performance of several baseline models alongside our MSME method trained on tabular meteorological data only, as depicted in Table IV. SVR performs unfavorably with a MAPE of 16.79%. In contrast, deep learning models such as the Transformer and 1D-CNN offer more competitive results, with MAPEs of 13.56% and 14.38%, respectively. LSTMs and their variants, including CNN-with-LSTM and FTA-LSTM (an LSTM combining both feature and temporal attention mechanisms), also perform robustly with CNN-with-LSTM outperforming all other baselines with a MAPE of 13.39%. Our MSME architecture, integral to our MIS-ME framework, achieves a MAPE of 15.26%, demonstrating its effectiveness for further exploration with multimodal soil moisture estimation strategies.

C. Performance Analysis of Image-Only Models

In evaluating image-only models for soil moisture estimation, we observe varied performance across several architectures as shown in Table IV. MobileNetV2 stands out with the lowest MAPE of 12.3%, likely due to its efficient design that balances complexity and performance. This makes it particularly effective for processing soil patch images. In comparison, ResNet18 shows decent results with a MAPE of 15.95%. EfficientNetV2 and InceptionV3 do not perform as well, recording higher MAPEs of 19.35% and 19.62%. Overall, the performance disparity among these models highlights the importance of choosing the right architecture based on the specific nature and scale of the data involved.

D. Performance Analysis of MIS-ME

1) MIS-ME with Multi-modal Concat: In examining the effectiveness of the Multi-modal Concat approach within our MIS-ME framework, significant performance improvements are observed when compared to the unimodal approaches. As outlined in Table IV, the Multi-modal Concat strategy using MobileNetV2 achieves a notable MAPE of 10.99% and outperforms all image-only and meteo-only models. This improvement underscores the impact of integrating both meteorological and image data, where the concatenated approach leverages information from both modalities, thereby resulting in more accurate and robust soil moisture predictions.

2) MIS-ME with Hybrid Loss: The hybrid loss approach within our MIS-ME framework emerges as the most effective model, achieving the lowest MAPE of 10.79% with the MobileNetV2 architecture, as shown in Table IV. This method finely balances the loss contributions from concatenated data ($\delta = 0.9$), meteorological data ($\gamma = 0.0$), and image data ($1 - \delta - \gamma = 0.1$), where the meteo loss component has no weightage, effectively prioritizing the concatenated and image losses. This configuration not only outperforms the unimodal approaches but also slightly improves over the Multi-modal Concat strategy. The strategic allocation of these weights ensures optimal utilization of both meteorological and image data, enhancing the predictive accuracy. Details on the tuning of these weighting coefficients are discussed in Section V-F3. This method highlights the significance of customized loss management in multi-modal learning frameworks, making the Hybrid Loss model the top pick of our three-way MIS-ME framework for soil moisture estimation.

3) MIS-ME with Learnable Parameter: The Learnable Parameter approach within our MIS-ME framework allows for dynamic weighting of meteorological and image-derived features, adapting the model's reliance on each data type based on their predictive value. Table IV shows that this approach achieves a MAPE of 11.16% with the MobileNetV2 architecture, which is competitive with the other multimodal approaches and significantly outperforms the unimodal models. Notably, after training to convergence, the learned weights usually stabilize at $\alpha = 0.65$ and $\beta = 0.04$, signifying the greater impact of meteorological data in our dataset. This differential weighting confirms the nuanced integration capability of the learnable parameter approach, effectively harnessing the

TABLE IV: Result Comparison of MIS-ME with Baselines

Approach	Model type	MAE	MAPE
Only Meteo Data	SVR [21]	0.044	16.79
	MSME	0.04	15.26
	1D-CNN [21]	0.042	14.38
	LSTM [33]	0.043	15.31
	CNN-with-LSTM [21]	0.038	13.39
	FTA-LSTM [34]	0.044	14.86
	Transformer [35]	0.038	13.56
Only Image Data	ResNet18 [29]	0.043	15.95
	MoblieNetV2 [31]	0.032	12.3
	EfficientNetV2 [32]	0.046	19.35
	InceptionV3 [30]	0.051	19.62
MIS-ME with Multimodal Concat	ResNet18	0.033	12.5
	MoblieNetV2	**0.029**	**10.99**
	EfficientNetV2	0.038	13.52
	InceptionV3	0.037	13.74
MIS-ME with Hybrid Loss	ResNet18	0.033	12.45
	MoblieNetV2	**0.029**	**10.79**
	EfficientNetV2	0.035	13.65
	InceptionV3	0.035	14.04
MIS-ME with Learnable Parameter	ResNet18	0.034	12.85
	MoblieNetV2	**0.029**	**11.16**
	EfficientNetV2	0.036	13.44
	InceptionV3	0.035	13.91

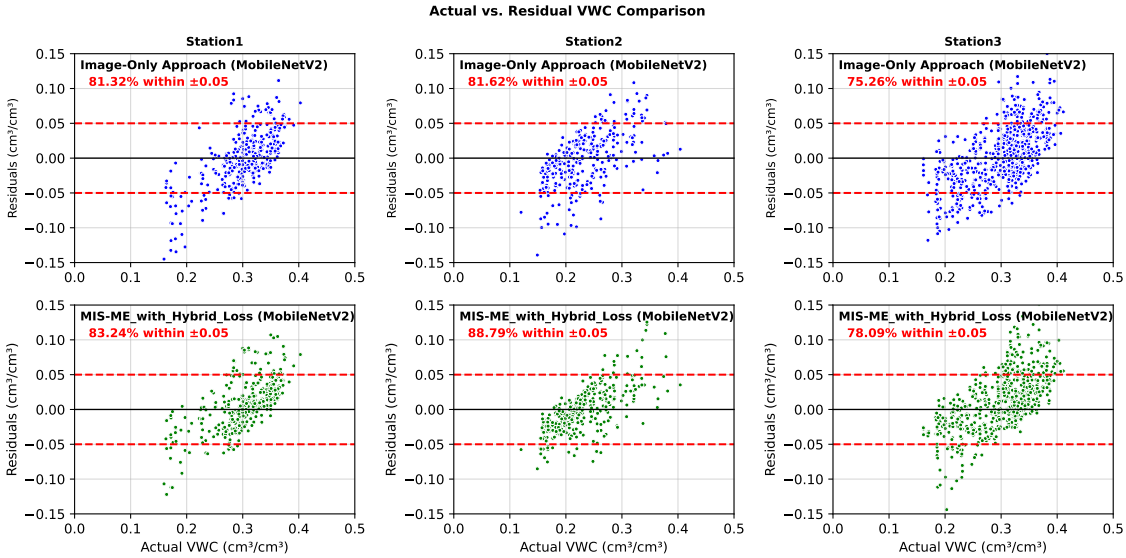

Fig. 8: Residual analysis of the Image-only approach and MIS-ME with Hybrid Loss across three stations for MobileNetV2. Higher percentage of residuals lie within the optimal range for MIS-ME than the regular image-only approach.

strengths of both modalities to enhance prediction accuracy.

Overall, the three-way MIS-ME framework outperforms all unimodal approaches convincingly, with the MobileNetV2 architecture coming out on top across all three approaches. Moreover, we see a significant improvement of more than 5% MAPE for EfficientV2 and InceptionV3 across all three approaches of the MIS-ME framework when compared to training these architectures solely using image data.

E. Station-wise Analysis of MIS-ME

1) Residual Analysis: Fig. 8 presents the residual analysis of the Image-only approach and MIS-ME with Hybrid Loss across three stations for MobileNetV2. The analysis shows that most residuals lie within the ideal [-0.05, 0.05] range. Notably, the MIS-ME-trained model demonstrates a higher concentration of residuals within this range than the Image-only approach. For instance, in the case of *Station1*, 83.24% of MIS-ME's residuals fall within this interval, in comparison to 81.32% for the image-only approach, reflecting a 2% enhancement. Similarly, for *Station2*, around 89% of MIS-ME's residuals are within the range versus 82% for MobileNetV2, a 7% improvement. For *Station3*, approximately 78% of MIS-ME's residuals are within the range, as opposed to 75% for MobileNetV2, marking a 3% improvement. These results indicate a tighter clustering of residuals near zero and a better model fit for the MIS-ME-trained model compared to the Image-only MobielnetV2 model.

2) Varying the Target Station Data: Fig. 9 illustrates the impact of incrementally introducing data from the target station to the training sample for all three approaches of our MIS-ME framework using MobileNetV2. This approach involves initially training the model using data exclusively from the non-target stations. For example, when targeting *Station1*, the model is trained solely with data from *Station2* and *Station3*, with no data from *Station1* included. This process is then progressively adjusted by adding 33.33%, 66.66%, and finally 100% of *Station1*'s training data, each time evaluating the model's performance on *Station1*'s test set. The results reveal that both *Station1* and *Station3* exhibit similar trends, with MAPE significantly reducing from around 18-16% to a stable range of 11-12% for all three approaches as more data from the target station is incorporated. In contrast, *Station2* shows a distinct pattern, starting with a high MAPE of around 36-40% when its data is initially absent during training, suggesting significantly different soil characteristics compared to the other stations. However, as *Station2* data is gradually introduced, the model's accuracy improves remarkably, stabilizing at a MAPE of 13-16%. This experiment shows that similar patterns

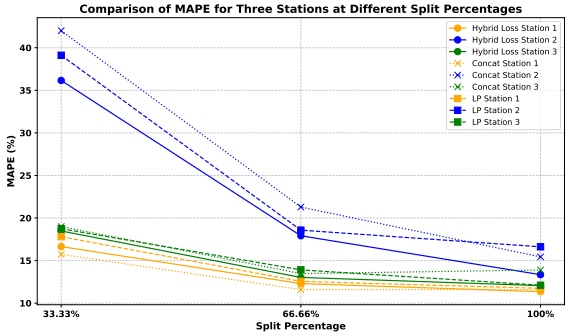

Fig. 9: Performance of all three approaches of the MIS-ME framework improves by increasing percentage (%) of the target station's data in the training sample across all stations. This showcases the requirement of geographic-specific features for a generalized model.

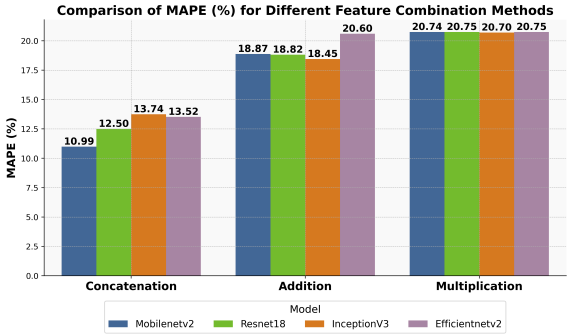

Fig. 10: Comparison of MAPE (%) for Different Feature Combination Methods

are required for optimal performance when predicting soil moisture in new locations.

F. Ablation Study

1) Feature Combination Methods for MIS-ME with Multimodal Concat: We explore the performance of different feature combination methods within the MIS-ME framework—specifically concatenation, addition, and multiplication. To prevent dimension mismatch for addition and multiplication, the image feature dimensions were downscaled to match those of meteorological features by passing them through two linear layers, each followed by batch normalization and ReLU activation, with a dropout layer incorporated after the first activation to reduce overfitting. As illustrated in Fig.10, direct concatenation demonstrates the best performance for all architectures. In contrast, addition and multiplication yield poorer results, with multiplication further lagging behind. These results show the effectiveness of concatenation in utilizing complementary information from both modalities, thus enhancing the model's predictive accuracy in soil moisture estimation. For this reason, we have gone with concatenation in our MIS-ME framework.

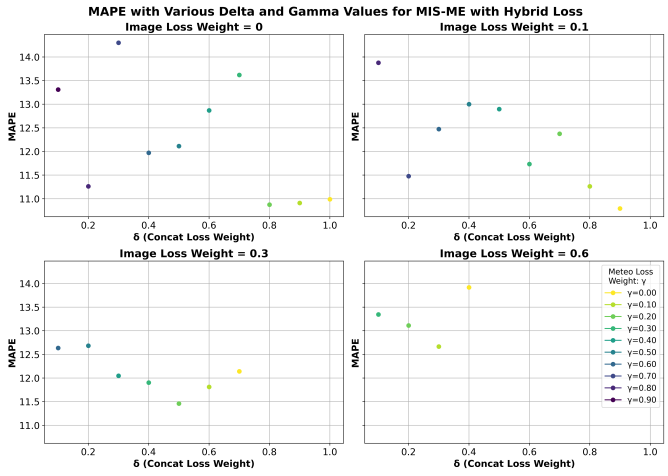

Fig. 11: Effect of Varying Hybrid Loss Coefficients δ and γ.

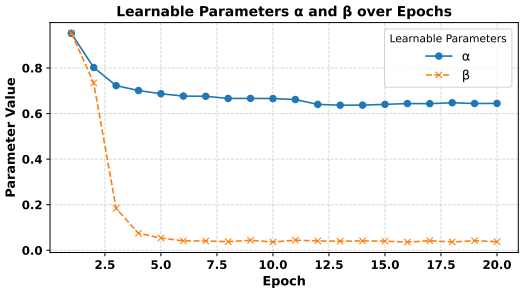

Fig. 12: Observing α and β of MIS-ME-with-Learnable-Parameters over Epochs

2) Optimization of Weighting Coefficients 'δ' & 'γ' in MIS-ME with Hybrid Loss: For optimizing the weighting coefficients in our MIS-ME with Hybrid Loss approach, we experimented with various combinations of δ, γ, and $1-\delta-\gamma$ to ensure different emphasis on each modality. As shown in Fig. 11, high weightage to concatenated loss coupled with low weightage to image loss or meteo loss generally yields better results.

The best performance is achieved with $\delta = 0.9$ and $1 - \delta - \gamma = 0.1$, resulting in a MAPE of 10.79% . This configuration prioritizes the concatenated and image losses while effectively excluding the meteorological loss, indicating that the combined information from these two losses is more reliable in accurate soil moisture prediction for our dataset. Interestingly, other combinations such as $\delta = 0.9$ and $\gamma = 0.1$ or $\delta = 0.8$ and $\gamma = 0.2$ also show competitive performance, with similar MAPEs. These findings highlight the significance of customizing the loss contributions to leverage the strengths of different data sources, thereby optimizing the predictive accuracy of our hybrid loss model.

3) Impact of Learnable Parameters on Prediction:

a) Analysis of Learned Weights in Enhancing Model Flexibility: Fig.12 illustrates the evolution of the learnable parameters α and β over the training epochs for the best-performing MobileNetV2 model. α and β dynamically adjust the contributions of the meteorological and image data, respectively, to obtain the final soil moisture prediction. Initially, both α and β start at 1.0 but gradually decrease as the model learns the optimal weights. By the end of training, α settles at approximately 0.65, and β at around 0.04. This indicates a higher reliance on meteorological data for accurate predictions in our dataset. Although this plot is specific to MobileNetV2, similar trends are observed across other architectures trained on our dataset. The dynamic adjustment of α and β enhances model flexibility by prioritizing the modality more relevant to soil moisture labels, thereby improving overall performance in soil moisture estimation.

b) Single vs. Dual Learnable Parameters: In our MIS-ME framework, we utilized two independent learnable parameters, α and β, for the meteorological and image modalities, respectively. To further explore the effectiveness of this approach, we conducted an ablation study using a single

learnable parameter, α, for the meteorological modality and $1 - \alpha$ for the image modality, forcing the model to learn a complementary relationship between them. However, results indicated that using such complementary parameters led to slightly higher MAPEs than independent learnable parameters. This suggests that allowing the model to independently adjust the weight of each modality offers better flexibility and performance for soil moisture estimation in our dataset.

VI. Conclusion

In this work, we explore machine-learning approaches with real-time images and tabular meteorological data collected from three ground stations for the soil moisture regression task. We highlight that this dataset is collected from the wild and resembles images taken from mobile phones. This makes them useful for evaluating real-time soil moisture estimation with commonly available images. We demonstrate that integrating image features with meteorological data using our proposed three-way MIS-ME framework significantly improves the performance of soil moisture regression compared to traditional approaches, which rely only on meteorological data. Notably, we experimented with feature combination techniques, hybrid loss functions, and learnable parameters. In the future, we aim to experiment with other methods, such as employing transfer-learning techniques like knowledge distillation. We conclude that including features of soil patches can positively impact the estimation of ground soil moisture, opening a new arena of research in the computational agriculture domain.

References

[1] T. E. Ochsner, M. H. Cosh, R. H. Cuenca, W. A. Dorigo, C. S. Draper, Y. Hagimoto, Y. H. Kerr, K. M. Larson, E. G. Njoku, E. E. Small, and M. Zreda, "State of the art in large-scale soil moisture monitoring," *Soil Science Society of America Journal*, vol. 77, no. 6, pp. 1888–1919, 2013.

[2] A. J. Phillips, N. K. Newlands, S. H. Liang, and B. H. Ellert, "Integrated sensing of soil moisture at the field-scale: Measuring, modeling and sharing for improved agricultural decision support," *Computers and Electronics in Agriculture*, vol. 107, pp. 73–88, 2014.

[3] X. Wu, M. Liu, and Y. Wu, "In-situ soil moisture sensing: Optimal sensor placement and field estimation," *ACM Trans. Sen. Netw.*, vol. 8, no. 4, sep 2012.

[4] Y. Feng, Y. Xie, D. Ganesan, and J. Xiong, "Lte-based low-cost and low-power soil moisture sensing," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '22. New York, NY, USA: ACM, 2023, p. 421–434.

[5] F. J. Pierce and P. Nowak, "Aspects of precision agriculture," ser. Advances in Agronomy, D. L. Sparks, Ed. Academic Press, 1999, vol. 67, pp. 1–85.

[6] S. A. Bhat and N.-F. Huang, "Big data and ai revolution in precision agriculture: Survey and challenges," *IEEE Access*, vol. 9, pp. 110 209–110 222, 2021.

[7] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2020.

[8] N. Filipović, S. Brdar, G. Mimić, O. Marko, and V. Crnojević, "Regional soil moisture prediction system based on long short-term memory network," *Biosystems engineering*, vol. 213, pp. 30–38, 2022.

[9] H. Shokati, M. Mashal, A. Noroozi, S. Mirzaei, and Z. Mohammadi-Doqozloo, "Assessing soil moisture levels using visible uav imagery and machine learning models," *Remote Sensing Applications: Society and Environment*, vol. 32, p. 101076, 2023.

[10] J. Yu, X. Zhang, L. Xu, J. Dong, and L. Zhangzhong, "A hybrid cnn-gru model for predicting soil moisture in maize root zone," *Agricultural Water Management*, vol. 245, p. 106649, 2021.

[11] R. Orth *et al.*, "Global soil moisture data derived through machine learning trained with in-situ measurements," *Scientific Data*, vol. 8, no. 1, pp. 1–14, 2021.

[12] A. Singh and K. Gaurav, "Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images," *Scientific Reports*, vol. 13, no. 1, p. 2251, 2023.

[13] L. Bertalan, I. Holb, A. Pataki, G. Négyesi, G. Szabó, A. K. Szalóki, and S. Szabó, "Uav-based multispectral and thermal cameras to predict soil water content–a machine learning approach," *Computers and Electronics in Agriculture*, vol. 200, p. 107262, 2022.

[14] A. Mansur, H. Abdullah, H. Syahputra, B. Benaissa, and F. Harahap, "An image processing techniques used for soil moisture inspection and classification," in *Proceedings of the 4th International Conference on Innovation in Education, Science and Culture, ICIESC 2022, 11 October 2022, Medan, Indonesia*, 2022.

[15] D. Kim, T. Kim, J. Jeon, and Y. Son, "Convolutional neural network-based soil water content and density prediction model for agricultural land using soil surface images," *Applied Sciences*, vol. 13, no. 5, p. 2936, 2023.

[16] A. S. Sagayaraj, S. Kabilesh, D. Mohanapriya, and A. Anandkumar, "Determination of soil moisture content using image processing-a survey," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2021, pp. 1101–1106.

[17] T. C. on Revision of Manual 70, "Evaporation, evapotranspiration, and irrigation water requirements." American Society of Civil Engineers, 2016.

[18] R. Fu, L. Xie, T. Liu, B. Zheng, Y. Zhang, and S. Hu, "A soil moisture prediction model, based on depth and water balance equation: A case study of the xilingol league grassland," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 1374, 2023.

[19] S. Prakash, A. Sharma, and S. S. Sahu, "Soil moisture prediction using machine learning," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, pp. 1–6.

[20] R. Togneri, D. F. dos Santos, G. Camponogara, H. Nagano, G. Custodio, R. Prati, S. Fernandes, and C. Kamienski, "Soil moisture forecast for smart irrigation: The primetime for machine learning," *Expert Systems with Applications*, vol. 207, p. 117653, 2022.

[21] Y. Wang, L. Shi, Y. Hu, X. Hu, W. Song, and L. Wang, "A comprehensive study of deep learning for soil moisture prediction," *Hydrology and Earth System Sciences Discussions*, vol. 2023, pp. 1–38, 2023.

[22] Y. Cai, W. Zheng, X. Zhang, L. Zhangzhong, and X. Xue, "Research on soil moisture prediction model based on deep learning," *PloS one*, vol. 14, no. 4, p. e0214508, 2019.

[23] Q. Li, Z. Li, W. Shangguan, X. Wang, L. Li, and F. Yu, "Improving soil moisture prediction using a novel encoder-decoder model with residual learning," *Computers and Electronics in Agriculture*, vol. 195, p. 106816, 2022.

[24] M. F. Celik, M. S. Isik, O. Yuzugullu, N. Fajraoui, and E. Erten, "Soil moisture prediction from remote sensing images coupled with climate, soil texture and topography via deep learning," *Remote Sensing*, vol. 14, no. 21, p. 5584, 2022.

[25] E. H. Hegazi, A. A. Samak, L. Yang, R. Huang, and J. Huang, "Prediction of soil moisture content from sentinel-2 images using convolutional neural network (cnn)," *Agronomy*, vol. 13, no. 3, p. 656, 2023.

[26] A. Habiboullah and M. A. Louly, "Soil moisture prediction using ndvi and nsmi satellite data: Vit-based models and convlstm-based model," *SN Computer Science*, vol. 4, no. 2, p. 140, 2023.

[27] R. Ding, H. Jin, D. Xiang, X. Wang, Y. Zhang, D. Shen, L. Su, W. Hao, M. Tao, X. Wang, and C. Zhou, "Soil moisture sensing with uav-mounted ir-uwb radar and deep learning," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 1, mar 2023.

[28] G. Jocher, "Yolov5 by ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016, pp. 770–778.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE CVPR*, 2016, pp. 2818–2826.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE CVPR*, 2018, pp. 4510–4520.

[32] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*. PMLR, 2021, pp. 10 096–10 106.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[34] Y. Ding, Y. Zhu, J. Feng, P. Zhang, and Z. Cheng, "Interpretable spatio-temporal attention lstm model for flood forecasting," *Neurocomputing*, vol. 403, pp. 348–359, 2020.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.