Advancing Precision Agriculture: Machine Learning-Enhanced GPR Analysis for Root-Zone Soil Moisture Assessment in Mega Farms

Himan Namdari, Majid Moradikia, Seyed Zekavat, *Senior Member, IEEE*, Radwin Askari, Oren Mangoubi, and Doug Petkie, *Senior Member, IEEE*

***Abstract*—In this article, we investigate an intelligent ground penetrating radar (GPR) that facilitates root-zone soil moisture estimation, a key parameter in precision agriculture. To create an intelligent GPR, we must train machine learning (ML) methods applied to the GPR-received signal. This process requires a large number of labeled GPR data that would be time-consuming and labor-intensive if created via field measurements. This article uses gprMAX software to emulate *drone-coupled GPR* received signal to generate large-scale data for training ML models. The data are created via a 1.5 GHz Ricker waveform considering a three-layer soil consistent with a realistic soil horizon model. The approach is structured as follows: first, we generate a synthetic dataset using gprMAX. Feature engineering techniques are then employed to extract meaningful components from the GPR signals, followed by a rigorous selection process to identify the most effective ML model for soil moisture prediction. Finally, we validate our model by integrating synthetic data with real GPR data collected at the *SoilX* lab at Worcester Polytechnic Institute, enhancing prediction accuracy and generalization capability. Our proposed model achieves an overall average root-mean-squared error of 0.5%, and 1.56 cm for moisture and depth estimations, respectively. The proposed intelligent GPR, when installed on a drone, enables high horizontal (e.g., 10 m) and vertical (e.g., 1.5 cm) resolution and high penetration depth (beyond 2 m) megafarm root-zone 3-D moisture map creation. Thus, it offers much higher capabilities when compared to traditional methods, such as synthetic aperture radar and satellite imaging. These results facilitate efficient farming practices, such as optimizing irrigation models, for better crop yields and food security.**

***Index Terms*—Feature extraction, gprMAX simulations, ground penetration radar (GPR), machine learning (ML), root zone soil moisture estimation (SME).**

Received 25 March 2024; revised 14 August 2024; accepted 29 August 2024. This work was supported by USDA under Grant NR223A750013G032. This article was recommended by Associate Editor Z. Liu. *(Corresponding author: Himan Namdari.)*

Himan Namdari, Majid Moradikia, Seyed Zekavat, and Oren Mangoubi are with the Data Science Program, Worcester Polytechnic Institute (WPI), Worcester, MA 01609 USA (e-mail: hnamdari@wpi.edu; majidemoradikiah@gmail.com; rezaz@wpi.edu; omanoubi@wpi.edu).

Doug Petkie is with the Physics Department, Worcester Polytechnic Institute (WPI), Worcester, MA 01609 USA (e-mail: dtpetkie@wpi.edu).

Radwin Askari is with the Department of Geological and Mining Engineering and Sciences, Michigan Technological University, Houghton, MI 48105 USA (e-mail: raskari@mtu.edu).

Digital Object Identifier 10.1109/TAFE.2024.3455238

I. Introduction

SUBSURFACE characterization plays a crucial role in diverse fields, such as agriculture, forestry, civil engineering, and space exploration [1], [2], [3], [4]. In the agricultural sector, understanding the spatial distribution of root-zone soil moisture is key to developing efficient irrigation systems, which are essential for the sustainability of large-scale farming operations. The efficient management of irrigation is paramount since, according to the Food and Agriculture Organization, agriculture accounts for 70% of global freshwater withdrawals [5]. Unfortunately, due to inefficient irrigation practices, about 60% of this water does not benefit the crops, leading to resource wastage and reduced agricultural productivity [6].

The quest for accurate soil moisture estimation (SME) has led to the exploration of numerous invasive and noninvasive methods, each with its own set of pros and cons [3], [4], [7], [8], [9], [10]. Invasive techniques involve direct soil sampling through probes for laboratory analysis [11], [12]. While offering high accuracy, these methods are criticized for their time-consuming and labor-intensive nature, disruption to plant and soil integrity, and lack of feasibility on the scale of megafarms [13]. On the other hand, non-invasive approaches, such as time domain reflectometry (TDR) [14], satellite imagery [15], electrical resistivity tomography (ERT) [16], synthetic aperture radar (SAR) [17], [18], and ground penetrating radar (GPR) [3], [4], [19], provide a less intrusive means of obtaining SME data, preserving soil and plant health. TDR is known for its precision and satellite methods for their extensive coverage, yet they fall short in penetration depth, being limited to surface analysis (1–5 cm) [20], which is insufficient for assessing the root zones of crops, such as soybeans and corn, ranging from 15 to 180 cm [21]. In this context, the advent of air-coupled GPR technology marks a significant breakthrough, offering rapid, efficient data acquisition and deeper soil penetration, making it an ideal solution for comprehensive SME in large-scale agricultural settings [22], see Table I.

GPR transmits an electromagnetic (EM) wave toward the ground and processes the backscattered signal from various soil sublayers [23]. This technology allows for the quick and accurate gathering of extensive data, which is then used to assess

TABLE I
COMPARISON OF NON-INVASIVE METHODS FOR SME IN MEGA FARMS

Method	Vertical Resolution	Horizontal Resolution	Penetration Depth	Mega-Farm Application
Drone-Coupled GPR	0.01–3 m	0.1–1 m	up to 40 m	Fast and Efficient [34], [35]
ERT	0.1–1 m	0.5–2 m	up to 50 m	Labor intensive [16], [36]
TDR	0.1–1 m	N/A (Primarily vertical)	up to 1 m	Labor intensive [37], [38]
SAR	1–10 m	1–10 m	Only Topsoil (1–10 cm)	Low Penetration [17], [18]
Satellite Imaging	1–100 m (Spatial)	1–100 m	Only Topsoil (1–10 cm)	Low Penetration [39], [40]

subsurface soil features, including moisture levels, texture, and layering [13], [24], [25]. Soil moisture content and composition contribute to a distinctive value of soil permittivity for a given soil sample [26]. *For ease in presentation, this article calls relative permittivity as permittivity*. This permittivity, represented by ϵ_r, quantifies the soil's capacity to propagate electric fields. Soil permittivity significantly influences the signal shape, delay, and attenuation received by GPR, resulting in variations in signal time and frequency domain [24], [27]. In SME using GPR, channel permittivity has the most evident impact on the received signal characteristic [26].

To analyze the variations and changes in signals received through soil, traditional signal processing techniques for GPR have been utilized. These techniques include time–frequency analyses, such as Fourier transforms [28] and wavelet transforms [29], which help in understanding how soil diversity and other environmental factors influence the signal. Despite the effectiveness of these signal processing techniques, they fail to deal with the variability and complexity of soil compositions across different fields [30]. In many cases, the collected GPR data is not *labeled* correctly, making the dataset nonrobust and impractical for training purposes. Addressing this gap, machine learning (ML) techniques have become increasingly utilized due to their ability to extract subtle patterns from complex and noisy datasets autonomously, thereby offering a better understanding of soil characteristics. Moreover, these ML models allow the generalization across various soil conditions and adapting to different environments [10], [31].

To effectively deploy ML models, it is imperative to construct a versatile and, more critically, accurately labeled GPR training dataset. This is critical to maintain the model prone to overfitting and underfitting, capable of performing effectively in real-world scenarios. To address the challenges of manually labeling GPR data for training ML models, a labor-intensive and impractical process in megafarm, researchers have turned to the synthetic generation of GPR data. This approach uses software-based simulation, such as gprMax [4], [13], [25], [32], [33]. These challenges include the creation of small-sized training sets [13], the absence of realistic simulation parameters [25], a lack of validation against real-world data, and a dependence on artificial constructs like simulated rebars to create recognizable features, such as hyperbolas in the datasets [33]. Such challenges can adversely affect ML models' training efficiency and overall performance in accurately predicting soil channel parameters.

Processing GPR-collected raw data (backscattered signals) is computationally intensive. Thus, extracting features from GPR raw data to reduce data dimensions is crucial [41]. In addition, working with the data features instead of raw data highlights key data attributes [23] and mitigates the impact of noise and clutter [42]. The methods outlined in [13], [25], and [32] commonly employ GPR raw data for training ML models. While these comprehensive approaches are practical, they often encounter increased model complexity and a higher risk of overfitting. Training ML models using raw data over a set of extracted features results in more accurate and robust training and testing results [43]. For instance, in [33], principal component analysis (PCA) [44] is introduced to extract the *eight most dominant components* and shows the applicability of feature extraction in improving the ML model accuracy for SME. Similar to other studies, they assume the existence of external objects and rely on analyzing hyperbolas to estimate soil moisture. This methodology does not apply to widespread SME applications at megafarms.

The successful application of ML using GPR data necessitates large-scale synthetic data generated by accurate simulations that emulate the real-world soil channel natural components and their diversity [11], [45], [46]. To achieve this goal, this paper uses the widely-used gprMax software [47]. gprMax is an open-source tool specifically developed to simulate electromagnetic wave propagation, solving Maxwell's equations in three dimensions through the finite-difference time-domain method [48]. This software emulates various soil and GPR signals by dynamically changing simulation parameters, such as signal type, frequency, moisture level, soil composition, and layer thickness. In our proposed model, the synthetically generated labeled data is calibrated via real-collected data to ensure full consistency of the synthetically generated data. Next, feature engineering and ML algorithms are established to estimate the root-zone moisture level. The proposed four-stage framework is shown in Fig. 1. The contributions of this article are as follows.

1) *gprMax-based synthetic labeled data creation:* We create a comprehensive synthetic dataset via gprMax software. To emulate real soil conditions accurately, we ensure compatibility with the realistic soil scenarios considering the common land model (CLM) [45], [49] and hydraulic model [50]. Our simulations involve configuring permittivity values, determining the number of subsurface layers, establishing thickness profiles, among other parameters, and subsequently converting permittivity values into moisture volumetric water content (VWC) using the Topp equation [51].
2) *Synthetic data validation and calibration via real collected data:* We validate synthetic data quality via real-world measurements collected at the SoilX lab site at the Worcester Polytechnic Institute (WPI), Worcester, MA, USA campus. The fidelity of the synthetic data is verified against real GPR datasets using mean absolute error (MAE), ensuring the practicality of the synthetic dataset.

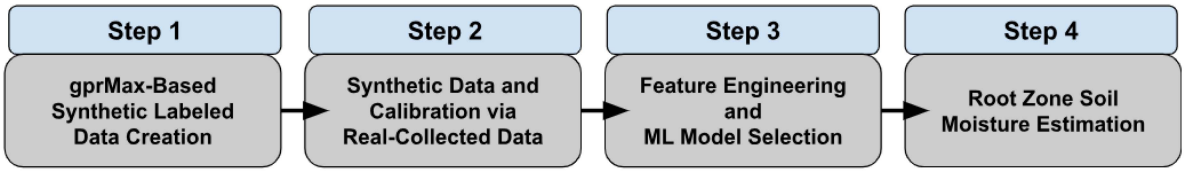


Fig. 1. Overview of the proposed Intelligent GPR-based SME framework, including four main interconnected steps.

3) *Feature engineering and ML model selection:* We employ feature engineering to extract and select significant features from the GPR data indicative of soil moisture content and subsurface layer thickness. Various ML models are then trained and evaluated to determine, which model best captures the complexities of the GPR data about SME. These include the peak extraction and statistical methods, such as feature importance using random forest (RF), analysis of variance (ANOVA), and PCA. Together, these methods prepare the data for our ML models, leading to better soil moisture predictions. To ensure these models work their best, we do hyperparameter tuning and check how well the models perform by comparing the training time, their root-mean-squared error (RMSE), and the coefficient of determination (R^2). These metrics help us understand how the models learn and how well they might perform on new data.
4) *Root Zone SME:* We will deploy the trained model on extracted feature sets, and combined dataset that includes synthetic and small-size real-world GPR data to predict the moisture content across different subsurface layers. The process involves validating and testing the model's performance, fine-tuning to adapt to the real data's complexities, and ultimately providing accurate, actionable soil moisture estimates at different depths.

To create an intelligent GPR, we need to implement and train ML methods applied to the received signal of a GPR system. This process requires a large number of labeled data. Creating such high-volume labeled data via real field measurements is time-consuming and labor-intensive. Thus, emulated labeled data that is consistent with real conditions is critical. This paper investigates the implementation of an intelligent GPR capable of estimating soil moisture via gprMax-based emulation.

The rest of this article is organized as follows. Section II introduces our proposed intelligent GPR-based framework for SME. Section III details the simulation procedures, including parameters and the specifics of our experimental setup, such as data collection methods and equipment. Section IV examines our findings and discusses their relevance and implications for our research. Finally, Section V concludes this article, which summarizes our main results and proposes avenues for future investigation.

II. Intelligent GPR-Based SME Framework

This section elaborates on each stage of the proposed Intelligent GPR-based SME Framework shown in Fig. 1.

TABLE II
Permittivity Conversion Based on Topp Equation (1)

relative permittivity	Moisture Content (%)	Moisture Type
5 – 9	9 – 18	Dry
10 – 16	19 – 28	Medium
17 – 21	29 – 35	Saturated

A. gprMax-Based Synthetic Labeled Dataset Creation

Root zone soil channel characteristics vary significantly depending on the geographic location and the type of crops grown, leading to a broad range of soil moisture profiles. Thus, to generate realistic synthetic GPR-labeled datasets, it is crucial to capture diverse conditions consistent with real-world scenarios by properly adopting simulated signal and soil channel parameters. Here, we aim to create HYDRO CLAMP algorithm1 that integrates two core channel models: Community Land Model [45] and Hydraulic Layer Model [46].

The community land model offers a detailed description of the soil layer structure, outlining essential layers, such as organic (0–10 cm), topsoil (10–60 cm), and subsoil (60–200 cm), see Fig. 2. The hydraulic layer model [46] reflects the increase in permittivity with soil moisture and depth. The HYDRO CLAMP algorithm begins by selecting permittivity values ranging from 5 to 21 for soil subsurface layers that are within 0–150 cm. This range is selected as it encompasses a broad spectrum of soil moisture conditions, from dry to fully saturated. For ease of labeling in our ML framework, we convert these permittivity values into VWC percentages using the Topp empirical equation [51]. The resulting VWC values span from 9% to 35%, as detailed in Table II. The Topp equation is as follows:

$$\epsilon_r = 3.03 + 9.3\theta + 146\theta^2 - 76.7\theta^3 \quad (1)$$

where ϵ represents the relative permittivity and θ the VWC. The algorithm also generates the depth range in 5–95 cm with a ten-step value increment. Top layer $L1$ is calculated by subtracting the combined depths of the second $(L2)$ and third $(L3)$ layers from the total depth, ensuring $D - L2 = L1$, and maintaining $L2 \geq L3$ for structural integrity. The parameters generated by Algorithm 1 serve as simulation input variables for the gprMax software [47]. This innovative approach enables us to generate synthetic data that accurately mirrors natural variations in soil moisture distribution and layer thickness across different regions.

The gprMax-based synthetic GPR data creation is based on a complex *propagation environment* that includes both the soil and the air channel, as well as the *corresponding received signals*. We define a 3-D propagation space with dimensions

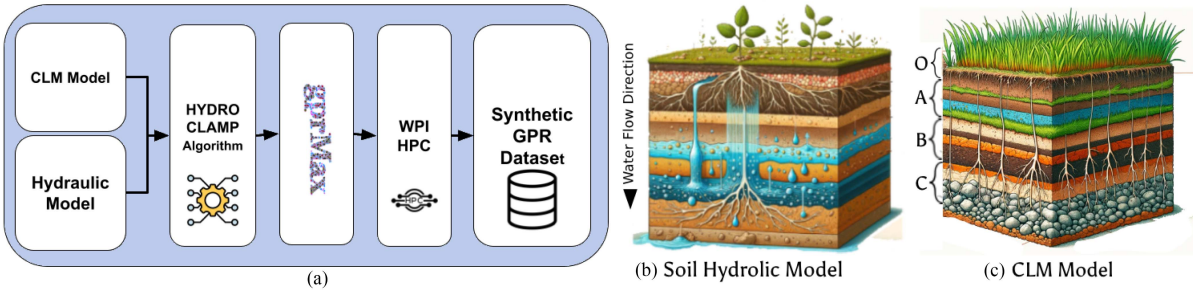


Fig. 2. (a) Proposed model for synthetic GPR dataset creation. (b) Soil Hydraulic model indicating water flow and layers distribution [50]. (c) CLM Model [45] indicating soil horizons O, A, B, C (Right) image sources [52].

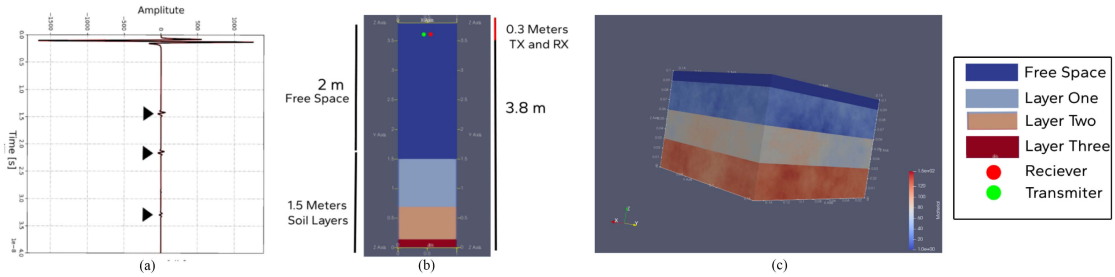


Fig. 3. (a) Received GPR signal (A-scan) with a central frequency of 1.5 GHz, propagation Z direction, and duration time of 40 ns, ▲ sign indicates layer reflections. (b) 2-D simulated environment, three-layer soil with a Tx and Rx above the surface. (c) Soil channel 3-D representation.

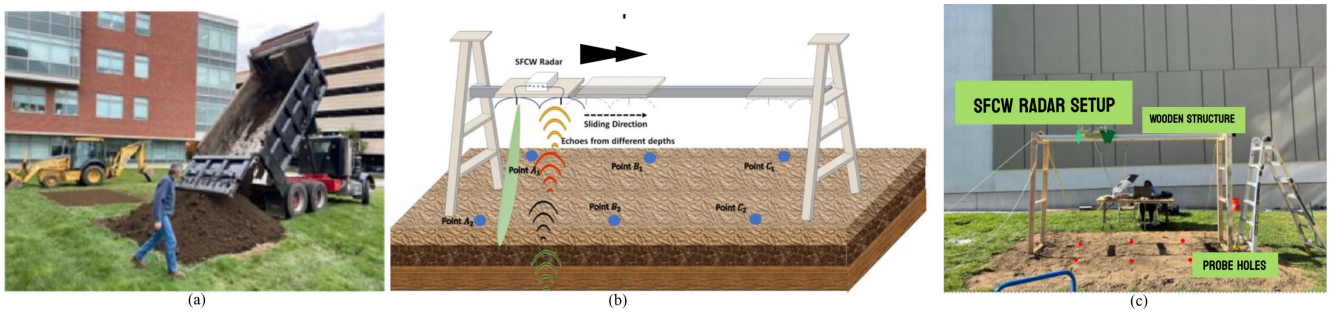
Algorithm 1: HYDRO CLAMP Algorithm.

Initialization:
Select relative permittivity values using the soil Hydraulic model II
Ensure generated permittivity follows $p1 < p2 < p3$
Permittivity to Soil Moisture:
Convert permittivities to moisture values using Topp's [51] equations:
    $m1, m2, m3 \leftarrow$ Function of$(p1, p2, p3)$
Pass $m1, m2, m3$ to gprMax simulator
Depth Value Generation:
Assume total soil channel model depth $D_{total} = 150$ cm
Generate depth values $d2$ and $d3$, in the range 5 to 95, in increments of 10 with constraints:
    $d2 > d3$
    $d1 = D_{total} - d2, d2 = D_{total} - (d1 + d3)$
Pass $d1, d2, d3$ to gprMax simulator
Simulation Preparation:
Insert depth and moisture values for gprMax input:
    InputString $\leftarrow (m1, d1) \| (m2, d2) \| (m3, d3)$
Write InputString to. in file for gprMax simulation

X, Y, Z, and a discretization cell size of d_s, selected to balance simulation accuracy and computational efficiency. The soil's conductivity σ S/m and its permeability μ_r were based on the default parameters for heterogeneous soil 0.01 and 1.0, respectively. Based on gprMax recommendations, the transmitter (Tx) and receiver (Rx) should be positioned at least 15 cells away from the domain's edge to minimize noise. We also specify a plate-perfect electric conductor material at the maximum depth of our simulation to reflect electromagnetic waves without any absorption or transmission shown in Fig. 3(b) and (c).

The impulse transmitted signal, operating at a central frequency of $f_0 = 1.5$ GHz and modeled with a Ricker waveform [53], propagates through the *Z*-axis in a duration of $t_0 = 40$ ns. This facilitates an in-depth analysis and modeling of electromagnetic wave interactions within these parameters. To simulate a drone surveying the ground from above, we place both the Tx and Rx at the center of the *X*-axis and the height of H meters, enabling the study of aerial GPR applications. The output from simulations is archived in `.out` files, which include the gprMax version, model title, iteration count, grid dimensions, time step, and source/Rx details. The Rx provides information about Rx positions and records time histories of electric and magnetic (EM) field components across three axes (Ex, Ey, Ez, Hx, Hy, and Hz). We used a postprocessing custom Python script extracting the vertical electric field component E_z component shown in Fig. 3(a).

Our initial benchmark study with single-layer soil simulations evaluates the impact of varying relative permittivity and the number of layers on the received signal. The large-scale simulations with three layers require substantial computational power, leading us to employ ten high-performance computing (HPC) nodes at WPI with ample RAM and multiple CPUs. Subsequently, simulation outputs were stored in. out files, while corresponding labels were stored in a separate text file, later consolidated into a CSV file. Each row in the CSV is a time-domain signal with its



(a) (b) (c)

Fig. 4. (a) Site preparation at WPI. (b) Proposed measurement campaign and wooden structure setup. (c) Actual setup, SFCW radar, Probe holes.

label appended at the end, facilitating the dataset for subsequent analysis or ML tasks.

B. Real and Synthetic Data Validation and Calibration

This step includes two main phases: real measurement campaign process and synthetic data validation and calibration.

1) Real Measurement Campaign and Considerations: The collection of our real-world data includes multiple phases: preparing the site, configuring the radar, arranging the measurement setup, and conducting the measurement campaign. For site preparation, we excavated a manually compacted area measuring $2 \times 3 \times 1.5$ m and filled it with loam, substituting the native soil to establish a controlled environment shown in Fig. 4. We used the AKELA AVMU radar, with a stepped-frequency continuous-wave (SFCW) [54] transceiver operating within 400–2000 MHz in 4096 steps. To simulate the drone-coupled radar movements, we built a wooden structure with the highest level of 2 m. We collected samples and recorded the ground truth moisture level using a probe at four depth points in our lab,[1] as shown in Fig. 4. Based on ground truth measurements, our soil composition typically consists of three primary layers defined based on moisture level.

2) Synthetic Data Validation and Calibration: Our real GPR data are a sequence of N coherent pulses transmitted, with the frequency of each successive pulse increasing by a fixed increment (0.39 GHz), denoted as Δf. The frequency of the nth pulse in the sequence can be expressed as : $f_n = f_0 + n\Delta f$, where f_0 is the starting frequency of the first pulse in the sequence, and n is the pulse index, ranging from 0 to $N-1$. This method allows the radar to effectively sweep through a range of frequencies, improving its resolution and detection capabilities by exploiting the frequency diversity of the reflected signals [54]. When a radar pulse is transmitted, its synchronous detector output is sampled, digitized, and stored. The output from in-phase (I) and quadrature (Q) channels forms a complex sample (real and imaginary components). These samples, termed "range bins," represent signals from a range window of length $\frac{c\tau}{2}$, where τ is the pulse width [54].

For an SFCW radar system, outputs from all range bins for N pulses in a burst are aggregated before processing. The initial step, "range binning" involves organizing this data into a range-pulse matrix, as depicted in Fig. 5, where each column represents a range bin corresponding to returns from N frequency-stepped pulses. By applying a discrete Fourier transform (DFT) to these columns, the radar subdivides the original range bin width, $\frac{c\tau}{2}$, into N smaller segments, where Δf is the frequency step, c is the speed of light, and τ is the pulsewidth. This process enhances the range resolution, yielding a high-resolution profile within each bin with subdivisions of width $\frac{c\tau}{2N}$, as per [54].

Signals	Pulses	Range Bins		
$RW_0 \times S_{f_0}$	P_1	S_{11}	S_{12}	S_{1M}
$RW_1 \times S_{(f_0+\Delta f)}$	P_2	S_{21}	S_{22}	$\cdots$
$RW_2 \times S_{(f_0+(2\Delta f))}$	P_3	S_{31}	S_{32}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$RW_n \times S_{(f_0+((N-1)\Delta f))}$	P_N	S_{N1}	S_{N2}	$\cdots S_{NM}$

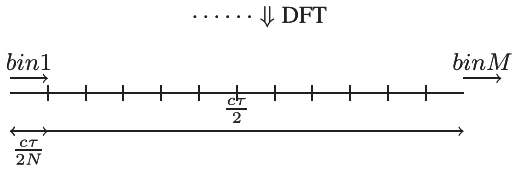


Fig. 5. Conversion of SFCW real-collected data by multiplying Ricker Wave R_W across each column.

To enhance the comparability between our real-world measurements obtained using the AKELA SFCW radar and the synthetic datasets generated via gprMax, which inherently employs Ricker waveforms for signal transmission, we introduce a novel postprocessing strategy in the Rx section of the SFCW radar system. Recognizing the distinct spectral characteristics imparted by the Ricker waveform, we applied a Ricker window in the frequency domain to the Fourier-transformed backscattered signals from each range bin. This operation effectively modulates the spectral content of the received SFCW radar data to mimic the frequency signature of the Ricker waveform used in the synthetic gprMax simulations. The rationale behind this approach is to filter the SFCW radar data in a manner that aligns its spectral characteristics with those of the gprMax-generated data, thereby facilitating a more fair comparison between the two datasets shown in Fig. 6. We then used MAE to compare the two signals; see Fig 6.

[1]https://soilx.wpi.edu/SoilX

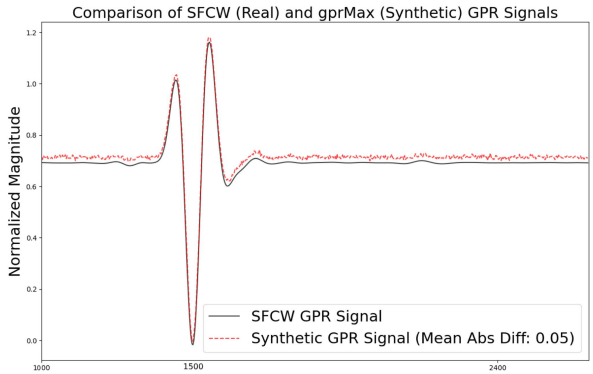


Fig. 6. Evaluation of processed SFCW and synthetic received signal.

C. Feature Engineering and ML Model Selection

Soil characteristics, such as permittivity, layer thickness, and the number of layers, directly impact the shape of the GPR time-domain signal. These factors influence the signal's amplitude, phase, and delay. High permittivity, for instance, slows wave speed due to increased electrical energy storage capacity. The feature extraction step is thus critical in identifying attributes like attenuation and time delays, which are pivotal in assessing soil conditions, such as moisture content, through the power delay profile. This process is essential for enhancing ML models' predictive accuracy.

The received signal is created as an interaction between the transmitted signal and the soil channel. Thus, the soil channel features are embedded in the received signal, including the soil channel specification. A signal can be represented in both time and frequency domains, and various features can represent a signal in both domains. For a time-domain signal, the statistics of peaks and valleys and their location in the time domain are considered proper features that describe the signal. We use characteristics of the GPR signal that act like a unique signature indicating soil properties. These properties appear as changes in the received signal, including how much it weakens and delays and how its shape and strength vary.

To capture these changes, we use simple statistics summarizing these variations and using them as distinct features. For example, we look at peak characteristics, such as the highest and lowest values, the total area under the curve, and the distance between peaks. Our feature selection process involves statistical measures, such as average, standard deviation, how much the signal peaks, and the distribution symmetry. We also use quartiles to handle signals that do not follow a normal distribution, which is familiar with mixed soils, enhancing the accuracy and reliability of our models. We analyze these features using RF and ANOVA to determine, which are most important for differentiating soil types and conditions. This method and comparison of the result with the PCA method as a ground truth simplifies our calculations and speeds up training while improving the accuracy of our models. It makes them more adaptable to different soil conditions and better at providing accurate GPR survey results. We have extracted these features and further processing steps via a customized Python code.

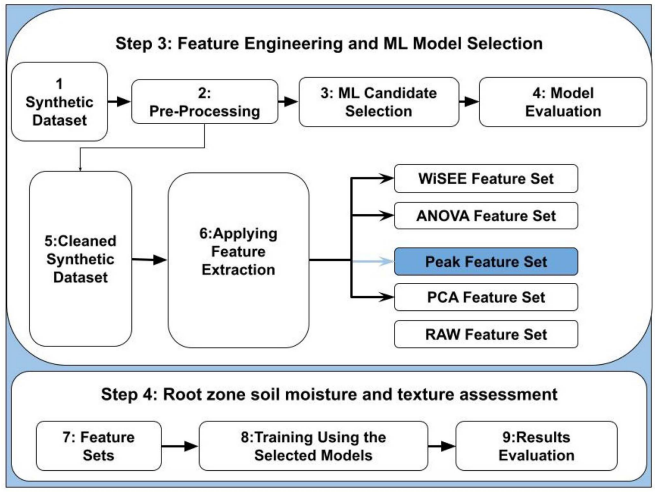


Fig. 7. ML model selection performance analysis by training raw data, feature extraction analysis using standard and innovative methods (step 2), performance analysis using selected ML and feature extraction method (step3).

The ML model selection step, shown in Fig. 7, is initiated by preprocessing to enhance the raw data quality and cleaning of our synthetic GPR dataset through normalization (Min–Max), automatic gain control, and zero time correction (Normal Move-out Correction). When choosing our ML model for analyzing the GPR dataset, we should consider our dataset characteristics, such as nonlinearity, high-dimensionality, and multiregression prediction problems. Therefore, we selected four main ML models for processing and evaluation: RF [55], gradient boosting regression (GBR) [56], support vector machine (SVM) [57], and neural networks (NN) [58] all capable of processing nonlinear and complex datasets. We employed hyperparameter tuning using grid search when trained using the preprocessed GPR dataset to optimize their performance and fair comparison. The effectiveness and performance of each model were quantified using RMSE coefficient of determination (R^2) and training time, offering insights into their goodness of fit and generalization capabilities in step two outlined in Fig. 7.

We integrated comprehensive data processing and model training phases into our framework to enhance our model selection, architecture, and parameter discussion. We selected different ML models; the RF model utilizes an ensemble of decision trees. Its architecture consists of 100 trees, and the final predictions were made by averaging the outputs of these trees. GBR builds trees sequentially, optimizing for errors of previous trees, making it powerful for handling nonlinear relationships. SVM employs a hyperplane to classify data points with a maximum margin, which is suitable for high-dimensional spaces and cases with a clear margin of separation. The architecture of SVM involves the kernel trick, which transforms input data into higher dimensions to find the optimal hyperplane. With its multilayered architecture, NN leverages neurons and activation functions to learn complex patterns from the data. The NN architecture is designed to handle the specific shape of the input data from the GPR system. There are two hidden layers, each half the size of the first layer of neurons using ReLU activation and an output

TABLE III
SIMULATION PARAMETERS OF GPR DATA

Parameter	Value
Simulation Domain ($X \times Y \times Z$)	$1.5 \times 1.5 \times 0.002$ m
Model Discretization size (d_s)	0.002 m
Signal Central frequency (f_0)	0.825 GHz
Signal Waveform	Ricker waveform
Signal Propagation axis	Z-axis
Signal duration (t_0)	40 ns
Signal Data points	8100
Total propagation time	40 ns
Simulation Labels	6 per layer
Labeled sample length	8106
Number of simulations	150 K
Raw dataset dimensions	$150K \times 8,106$

TABLE IV
ML MODELS PERFORMANCE EVALUATION USING RAW GPR DATA

Model	Train RMSE	Test RMSE	Training Time	R^2 Value
RF	**5.11**	**5.21**	**33(s)**	**0.97**
GBR	6.32	7.45	130(s)	0.89
SVR	7.39	9.56	58(s)	0.85
NN	5.28	5.30	62(s)	0.92

TABLE V
ACTUAL VERSUS PREDICTED LABELS AVERAGE RMSE ERROR USING RF MODEL AND PEAK FEATURE SET

Parameters	Actual	Predicted	Error	RMSE by Type	RMSE
M_1 (%)	6	6.39	0.39	0.5 (%)	1.1
M_2 (%)	13	13.78	0.78		
M_3 (%)	34	34.39	0.39		
D_1 (cm)	60	61.17	1.17	1.56 (cm)	
D_2 (cm)	70	71.56	1.56		
D_3 (cm)	20	21.95	1.95		

TABLE VI
GPR DIMENSION REDUCTION USING FEATURE EXTRACTION METHODS

Dataset	A-scan Size	Reduction Rate (%)
Raw GPR	8000	-
ANOVA Feature Set	100	98.75
DS Feature Set	5	99.93
Peak Feature Set	5	99.93
PCA Feature Set	2	99.97

layer with six neurons corresponding to the moisture and depth values using a linear activation. Each model's training phase involved hyperparameter tuning and cross-validation to ensure optimal performance and generalization. Our data processing pipeline included normalization, feature selection, and handling missing values, which were crucial steps to enhance model accuracy and efficiency.

Notably, recent research has utilized raw GPR data for ML model training [13], [25]. However, such data includes unprocessed and noisy signals that cause overfitting. This is because models might learn the noise as the signal features, interfering with their generalization capabilities on unseen datasets. Consequently, applying feature selection and extraction techniques becomes crucial and plays a pivotal role in mitigating noise, reducing data dimensionality outlined in Table VI, thus significantly enhancing the models' ability to learn and interpret the data effectively. This underscores the vital need for feature extraction strategies in overcoming the innate obstacles presented by GPR data analysis [59].

For feature extraction analysis, we explore the impacts of changing permittivity and layer thickness on the received signal in our single-layer generated dataset. We randomly change the layer's permittivity and thickness and analyze the received signal shape, number of reflections, and attenuation channels. Based on our observation, the proposed approach includes four methods of extracting important time domain features using 1) descriptive statistics (DS) proposed in our recent paper [3], 2) ANOVA [60], 3) peak feature extraction and engineering approach, and 4) PCA [44] as shown in Fig. 7. In particular, we have utilized descriptive statistical values, such as minimum, maximum, mean, standard deviation, quartiles, kurtosis, and skewness. To further improve the quality of the attributes extracted from the signal, we employed an RF model to refine the feature set obtained through these statistical measures, considering their highest importance. The minimum and maximum values illuminate the signal's reflection range. At the same time, the mean, standard deviation, and quartiles reveal average signal strength, variability, and distribution, which are crucial for identifying soil conditions. Skewness and kurtosis refine this analysis by indicating asymmetry and extreme values, pinpointing changes in soil composition or moisture. This collective analysis is key for accurately extracting patterns and understanding soil scenarios through GPR signals.

ANOVA can assess feature importance in datasets with many attributes by comparing variations within and between feature groups to the overall variation in the label [61]. Using ANOVA, we evaluated the significance of each attribute in our dataset, which consists of 8100 attributes per sample across six distinct labels. This statistical method allows us to explore how variations within and between groups of features contribute to the overall variability observed in each label. The F-statistic, derived from ANOVA, is central to this analysis, as it quantifies the discrepancy in mean values across different feature groups relative to the overall mean. A low p-value accompanying the F-statistic indicates a significant difference in group means, suggesting that these variations are not merely due to chance. Therefore, we conducted individual ANOVA tests for each label to determine each attribute's relative importance. We then compiled and compared these findings to rank them based on F-statistic value in order of significance, enabling us to pinpoint the top 100 attributes that consistently showed the highest impact.

The peak extraction method collects the peak-related information from GPR signals. These features are peak' count, amplitude, spacing, area under peaks, and their indexes in time. These features each help to identify attributes of measuring soil layer thickness and consistency, with the area under peaks indicating energy linked to soil features. The peak time index is essential for estimating depth. This approach simplifies mapping soil moisture and structure, providing a clearer picture of the soil's composition demonstrated in Fig. 7. The primary goal is to extract relevant, useful signal attributes highly related to the impact of moisture and layer variation. Then, RF will be used to refine these values and pick the most important features.

We evaluated the effectiveness of different feature extraction methods by training the selected ML model (RF and NN) using our extracted feature sets. Then, we compared their result with the trained model using raw and PCA feature sets separately, the results shown in Table IV.

We further evaluated and selected ML to process the generated data efficiently, demonstrating a good fit for the GPR data by training several models. Models including RF [55], GBR [56], SVM [57], and NN [58] were selected considering multiregression problem and nonlinearity in our GPR. We employed hyperparameter tuning using grid search [62] to optimize them. The model estimation accuracy was evaluated using the training time (s), average RMSE, and coefficient of determination (R^2) for learning efficiency and generalizability.

D. Root Zone SME

In the validation phase for our synthetic dataset, we adapted our real dataset, collected at WPI, to replicate the conditions simulated by gprMax. This crucial step ensures the synthetic data's fidelity, establishing a reliable baseline for subsequent model training and testing. Following this, our validation efforts for the ML-trained model centered on its performance using feature sets and merged datasets of synthetic and real-world GPR data to predict moisture content across various subsurface layers. The process begins with meticulous data preparation, segmenting the dataset into training, validation, and testing sets with ratios of 70%, 15%, and 15%, respectively. During the validation phase, we employ strategies, such as cross-validation to enhance the model's generalizability and perform hyperparameter tuning to optimize its performance.

Subsequently, the model undergoes rigorous testing with an unseen dataset to measure its generalization capability to new data. We employ RMSE as a metric for evaluating predictive accuracy while considering training time to assess the model's efficiency and performance. Our ultimate objective is to deliver precise, actionable soil moisture estimates at various depths, ensuring the model's reliability and practical applicability.

III. Experimental Setup

We designed a process to emulate labeled datasets using gprMax, encompassing diverse signal characteristics and soil parameter combinations. Each 3-D emulation dynamically adjusts soil layers and their moisture content using the parameters generated by the HYDRO CLAMP algorithm, ensuring realistic scenarios and precise control over soil parameters. To ensure data integrity, we kept the soil environment and the radar setup consistent across all simulations, effectively reducing noise displayed in Fig 3.

The 3-D simulation environment was set to 3.8 $\times$ 1.0 $\times$ 0.002 m and a discretization size of $d_s = 0.002$. Tx and Rx were positioned 2 m above the ground, mimicking drone-mounted GPR, with a 0.08-m separation [47]. To minimize the impact of signal reflections on the simulation's accuracy, the chosen default setting for the absorbing boundary conditions employs first-order complex frequency shifted perfectly matched layers. These layers are implemented with a uniform thickness of 10 cells along all six simulation domain boundaries, recommended for optimal absorption [47].

Our impulse transmitted signal, operating at a central frequency of $f_0 = 1.5$ GHz and modeled with a Ricker waveform [53], propagates through the *Z*-axis in a duration of $t_0 = 40$ ns. The received signal was discretized into a vector comprising 8100 data points, each representing the signal magnitude at discrete intervals over a total propagation time of 40 ns see Table III. This signal corresponds to six labels, including m_1, m_2, m_3 moisture levels and d_1, d_2, d_3 the associated depths. By running 150 000 such simulations, a large dataset with dimensions of 150 000 by 81,06 is obtained to support robust ML analysis and model training, shown in Fig 3(a). We used deletion and capping to handle missing values and outliers in our dataset. Missing values were addressed by removing rows with missing data, ensuring model training on complete information. Outliers were managed through capping, limiting extreme values to predefined upper and lower boundaries, thus reducing their impact on the model while preserving data integrity. The received signals are saved in an HDF5 format, efficiently organizing large numerical datasets. The Rx set captures the time history of both the electric field components (Ex, Ey, and Ez) and the magnetic field components (Hx, Hy, and Hz) at the Rx position. Custom Python scripts extracted the E_z component from `.out` files and to automate the adjustment of `.in` files, documenting simulation parameters. Data are stored in CSV format, and our small-scale simulations were performed on a Dell Latitude E7440, featuring a Core i7 processor and 32 GB RAM. In contrast, our more complex approach to simulating 150 000 simulations required a HPC setup with ten nodes, each endowed with 16 GB RAM and seven cores. The setup utilized the message-passing interface for parallel processing. This labeled training data supported optimal analyses, including training ML models and feature extraction.

Calibration of the synthetic soil with real channel involves replicating the AKELA GPR measurement setup at the WPI field. This includes radar height and soil moisture (measured by our moisture probe up to 1-m depth). Using the Topp equation [51], soil moisture has a key impact on soil permittivity. Thus, we ensured that permittivity values and layer depth accurately reflected actual soil moisture. Our experiment validated the emulated data against real data obtained from AKILA radar. This radar generates SFCW waveforms over 4096 frequency components ranging from 400 MHz to 2 GHz, captured at WPI grounds. Our emulated data are created via Ricker waveform. To validate the emulated results generated by the Ricker waveform against real data, we mapped the SFCW waveform into the Ricker waveform. Then, we assess the similarity between the mapped SFCW and synthetically generated data. The mapped data array and synthetic data array frequency ranges should be properly before similarity evaluation. Fig. 8 represents the details of this process. The mapping was implemented by applying a normalized mask derived from the Fourier transform of the Ricker waveform to the SFCW frequency domain data. The measure for similarity is mean Euclidean distance, also called mean square error. Preliminary results show a promising similarity between real and synthetic datasets, indicating effective replication of field conditions.

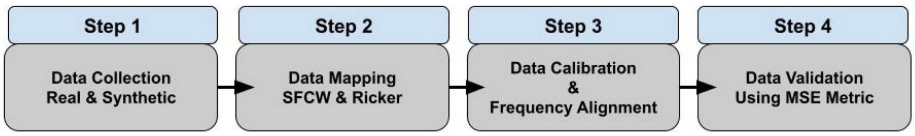


Fig. 8. Overview of the proposed calibration method of synthetic datasets against real-world data.

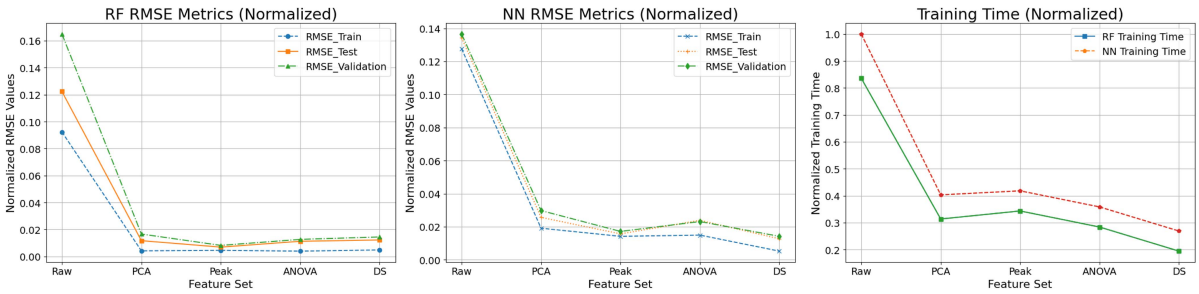


Fig. 9. Left and middle figures show the average normalized RMSE for training, testing, and validation phases of two ML models when trained on different extracted future sets (x-axis), and the right plot showcases the normalized training times, contrasting the computational efficiency in RF and NN, across various feature sets. In the RF and NN result plots (left and middle), the RF model demonstrated the best performance with the lowest RMSE. In addition, the DS feature set required the least training time for the RF model.

ML training with raw data results identified the RF and NN models as top performers. Our study configured the RF to include 1000 estimators, ensuring a robust ensemble model. Each tree was allowed to grow to a maximum depth of 3, with node splitting guided by the mean squared error criterion to minimize prediction variance. Bootstrap aggregation was employed to enhance model stability and mitigate overfitting. The NN incorporated a sequence of three dense layers, with the first two layers comprising 64 neurons each. These utilized the ReLU activation function to provide the model with the nonlinear learning capability. The final layer, responsible for outputting continuous values, contained six neurons with a linear activation function to match the regression task. Optimization during training was carried out using the `adam` algorithm, with a learning rate set to 0.001, and to prevent overfitting, an $L2$ regularization factor of 0.01 was applied. Model training was conducted over 100 epochs with a batch size of 16. A grid search method was utilized to fine-tune hyperparameters, ensuring that the chosen configuration offered the best predictive performance indicated by the lowest RMSE on the validation set, validation result shown in Fig. 9.

The labeled data was used in the training phase, leveraging sci-kit-learn and other Python libraries for preprocessing and ML training. Our training experiments, conducted on HPC shell and partially on `Google Colab`, the results of experiments confirmed that by eliminating low-importance features from the raw data, we could significantly improve the accuracy of training and the training time of ML models utilizing GPR data. Furthermore, to enrich our training dataset and improve the training accuracy, we combined the collected samples from the WPI site and integrated them with the synthetic dataset in the same format, shown in Fig. 4. This combined dataset was then used to train the selected models (RF and NN), and its effectiveness was assessed by evaluating the model's performance using the updated loss function as follows:

$$L = \frac{\sum_{i \in D_{\text{real}}} w_{\text{real}} \cdot l(y_i, \hat{y}_i) + \sum_{j \in D_{\text{synth}}} w_{\text{synth}} \cdot l(y_j, \hat{y}_j)}{|D_{\text{real}}| \cdot w_{\text{real}} + |D_{\text{synth}}| \cdot w_{\text{synth}}} \quad (2)$$

where D_{real}, D_{synth}, w_{real}, and w_{synth} are sets of indices and weights assigned to data points, respectively. In addition, $l(y, \hat{y})$ denotes the loss function that measures the difference between true y and predicted labels [63]. After aligning the processed real data (30 samples) with our synthetic format, we employed the aforementioned weighted training strategy by modifying the training loss function and concatenating the data for RF model training. The results showed a modest improvement, underscoring the value of incorporating real data into the training process. Despite the inefficiency of collecting real data, this method enable calibration of the trained models.

IV. Result and Discussion

To properly compare the actual GPR signals obtained from SFCW methods against synthetic equivalents, we standardized the format of the real-world data to match that of the synthetic dataset. This ensures consistency in the evaluation process and subsequent analyses. We employed the MAE as our comparison metric to assess the similarity between the real and synthetic datasets. The analysis revealed an MAE of 0.05, indicating a minimal average discrepancy between the real and synthetic shown in Fig. 6.

The results of selecting the best model indicated that RF and NN models trained by raw GPR data maintained the lowest RMSE, achieved the lowest RMSE, with average training RMSEs of 5.1 and 5.21, and test RMSEs of 5.28 and 5.30, respectively. In addition, the training durations for the RF and

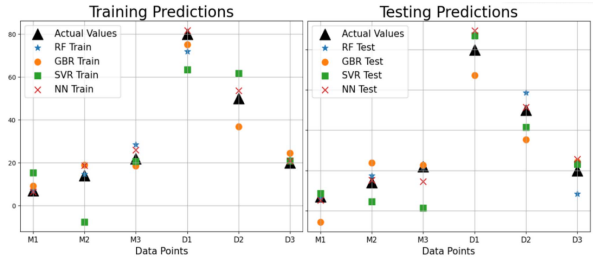


Fig. 10. Individual label comparative performance of RF and NN models.

NN models were 33 and 62 s, respectively, as shown in Table IV. The R^2 value indicates a robust fit to the dataset, underscoring the RF model's superior predictive accuracy and capability to capture the variance within the data effectively. The $R2$ value in the RF model achieved the highest value of 0.97, respecting 0.93 in the NN model depicted in Table IV. The GBR and SVR models, while showing acceptable RMSE values, have lower R^2 and are notably slower in training, suggesting a tradeoff between accuracy and speed as shown in Table IV.

In evaluating individual label predictions for a single sample, the RF model demonstrated higher training accuracy in predicting labels m_1, m_2, d_2, d_3, m_3, and d_1. Conversely, during the training phase, the NN model showed superior prediction accuracy for labels m_3 and d_1. When examining test accuracy, the NN model outperformed the RF model specifically for labels d_2 and d_3 . Both models exhibited comparable performance in the test phase for all other labels. Our documentation details these comparative results in Fig. 10 .

The result of the RF model with a peak feature set to predict various soil parameters focuses on moisture content and depth across three distinct layers. The predictions closely aligned with the actual values for moisture content, exhibiting minimal error rates across the three layers: 0.39% for layers 1 and 3 and a slightly higher 0.78% for layer 2 outlined in Table V. The depth predictions for each layer demonstrated greater variability, with errors of 1.17, 1.56, and 1.95 cm for layers 1, 2, and 3, respectively. Despite these variations, the overall performance of the RF model achieved an average RMSE of 1.1. This indicates a strong predictive capability, particularly for applications requiring precision in moisture content and soil depth measurements shown in Table V.

Applying various feature extractions resulted in a reduction in dimensionality for our GPR dataset. Notably, the ANOVA, DS, and Peak feature sets achieved reductions of over 98%, with the PCA feature set demonstrating an exceptional reduction rate, as shown in Table VI.

Our feature extraction result section evaluated the performance of RF and NN models trained with various feature sets. For both models, the RMSE for training, testing, and validation decreases significantly when moving from raw data to processed data through PCA, peak feature set, and ANOVA, with the lowest RMSE observed in the peak feature set. This trend suggests that data simplification and feature selection enhance model performance. The RF model consistently requires less time than the NN model across all feature sets in the training time plot, indicating a more efficient training process shown in Fig. 9.

Our comprehensive analysis, presented in Table VII, compared the performance of the RF and NN models on synthetic and integrated datasets. The integration of synthetic and real GPR data was achieved by first processing the received SFCW signal and then replicating these conditions in a gprMax simulation, and finally, we concatenated these real and synthetic datasets to train our model, leveraging the combined data to enhance the accuracy and robustness of our predictions.

The RF model improved with the integrated data, exhibiting a training RMSE of 0.89, a test RMSE of 0.92, and a validation RMSE of 1.10, while the training time slightly increased to 39 s. Conversely, the NN model displayed higher RMSE values for the integrated data, with a training RMSE of 1.8, a test RMSE of 1.9, and a validation RMSE of 2.2. The integrated data sets, with weighted training, slightly impacted both ML models' learning outcomes, indicating more efficient and precise performance when utilizing actual data.

In addition, we compared our proposed model with recently published methods identified in our literature review. Our model, which utilizes peak-based feature extraction, demonstrated superior accuracy and training time in estimating soil moisture content. Moreover, it provided precise predictions of soil depth and performance metrics, which the other approaches overlooked. As shown in Table VIII, the average error for predicting moisture is overall 0.52. This analysis indicated an average percentage error of approximately 4.88% using the Topp model to convert the error to the moisture error value, showcasing the expected variance in VWC due to errors in permittivity prediction. Furthermore, the average RMSE for depth prediction is 1.56 cm, highlighting the model's precision in estimating depth values. These predictions enable the reconstruction of a 3-D environment in our simulation platform. By aggregating these results, a comprehensive 3-D map of soil moisture and a profile map can be generated with the accuracy of our simulations and a close replica of the actual sampling points.

In Table VIII, we conducted a comparative analysis with recent approaches others, acknowledging differences in metrics while emphasizing the significance of validation set results and training time, aspects often overlooked or unreported in other studies. This comparison underscores the importance of realistic simulation, considering multiple evaluation metrics. The proposed moisture assessment approach will support mega-farm owners in creating an optimized irrigation model by making proper decisions on when, where, and how much water to apply. An intelligent GPR will archive this to facilitate soil moisture prediction and the creation of 3-D soil moisture maps. The historical 3-D maps collected over a period streamline root zone soil moisture forecasting, enabling farmers to optimally schedule the irrigation plans of megafarms. Thus, GPR technology combined with ML can significantly enhance precision agriculture practices, promoting efficient water usage and improved crop yields. Combining soil moisture measures with methodological, spatial, and irrigation history will be beneficial in improving the results.

TABLE VII
MODEL PERFORMANCE COMPARISON: SYNTHETIC DATA VERSUS INTEGRATED DATA

Model	Synthetic Data				Integrated Data (real & synthetic)			
	RMSE Train	RMSE Test	RMSE Validation	Training Time	RMSE Train	RMSE Test	RMSE Validation	Training Time
RF	**0.91**	**0.93**	**1.21**	**38**	**0.89**	**0.92**	**1.1**	**39s**
NN	1.9	2.1	2.3	56	1.8	1.9	2.2	61s

TABLE VIII
RESULT COMPARISON OF RECENT APPROACHES AND OUR PROPOSED MODEL

Article	ML Model	Data Type	Training Acc	Testing Acc	Validation Acc	Training Time
Barkataki et. al [32]	ANN	Synthetic	N/A	MAPE = 1.23	N/A	N/A
Giannakis et. al [33]	NN	Synthetic & Real	N/A	12% (SWC)	N/A	N/A
Our Proposed Model (SoilX)	**RF** & NN	Synthetic & Real	**RMSE = 0.91**	**RMSE = 0.93**	**RMSE =1.1**	**38 s**

V. CONCLUSION

This study develops a novel framework that integrates GPR with ML to estimate root-zone soil moisture and subsurface depths, significantly surpassing traditional methods. Our approach, particularly through RF and NN, demonstrates better accuracy in soil moisture and the associated depth predictions by employing a synthetic GPR, innovative feature extraction methods, and dataset enrichment using our real-world collected data. The broader implications of our research are substantial, providing a precise tool for soil moisture mapping that is vital for designing irrigation strategies, conserving water, and promoting soil health. Consequently, this contributes to advancing sustainable agricultural practices and food security. The versatility of our framework across various soil conditions highlights its potential to support sustainable farming and efficient water management globally. Looking ahead, we aim to redesign our intelligent GPR model considering diverse, realistic soil texture, and composition. This will increase the temporal soil moisture assessment accuracy, critical to precision irrigation. Extensive field validations across various agricultural settings will be essential to gauge our framework's real-world applicability and impact. The intelligent method introduced in this article can also be used for soil nutrition management and pest control [64]. In addition, we also aim to explore other environmental information, such as humidity, temperature, soil texture, soil composition, air pressure, and sensory information, such as satellite, IR, and optical imaging, which can be integrated with GPR data through a multimodal ML technique to increase the accuracy of the proposed approach. Accurate soil moisture and depth estimations, combined with meteorological data and satellite imagery, can be applied to multimodal ML models. This integration produces multidisciplinary insights, improving agricultural and subsurface assessments.

REFERENCES

[1] N. J. R. Fernandez, J. M. Sabater, P. Richaume, A. A. Yaari, and Y. H. Kerr, "SMOS soil moisture retrieval over grasslands: SMOS L2 algorithm or a new neural network?," *Geophysical Res. Lett.*, 2015. [Online]. Available: https://doi.org/10.1002/2015GL063388

[2] D. Zhang and G. Zhou, "Estimation of soil moisture from optical and thermal remote sensing: A review," *Sensors*, vol. 16, no. 8, 2016, Art. no. 1308.

[3] H. Namdari, M. Moradikia, D. T. Petkie, R. Askari, and S. Zekavat, "Comprehensive GPR signal analysis via descriptive statistics and machine learning," in *Proc. IEEE Int. Conf. Wireless Space Extreme Environ.*, 2023, pp. 127–132.

[4] V. Filardi et al., "Data-driven soil water content estimation at multiple depths using SFCW GPR," in *Proc. IEEE Int. Opportunity Res. Scholars Symp.*, 2023, pp. 86–90.

[5] FAO, "Coping with water scarcity—An action framework for agriculture and food security," 2012. [Online]. Available: http://www.fao.org/3/i3015e/i3015e.pdf

[6] J. Wallace, "Increasing agricultural water use efficiency to meet future food production," *Agriculture Ecosystems Environ.*, vol. 82, no. 1-3, pp. 105–119, 2000.

[7] K. C. Kornelsen and P. Coulibaly, "Root-zone soil moisture estimation using data-driven methods," *Water Resour. Res.*, vol. 50, no. 4, pp. 2946–2962, 2014.

[8] V. Komarov, S. Wang, and J. Tang, "Permittivity and measurements," 2005.

[9] I. Rodriguez-Iturbe, P. D'odorico, A. Porporato, and L. Ridolfi, "On the spatial and temporal links between vegetation, climate, and soil moisture," *Water Resour. Res.*, vol. 35, no. 12, pp. 3709–3722, 1999.

[10] M. Rasol et al., "GPR monitoring for road transport infrastructure: A systematic review and machine learning insights," *Construction Building Mater.*, vol. 324, 2022, Art. no. 126686.

[11] C. Albergel et al., "From near-surface to root-zone soil moisture using an exponential filter: An assessment of the method based on in-situ observations and model simulations," *Hydrol. Earth Syst. Sci.*, vol. 12, no. 6, pp. 1323–1337, 2008.

[12] A. M. Peterson, W. D. Helgason, and A. M. Ireson, "Estimating field-scale root zone soil moisture using the cosmic-ray neutron probe," *Hydrol. Earth Syst. Sci.*, vol. 20, no. 4, pp. 1373–1385, 2016.

[13] N. Barkataki, S. Mazumdar, P. B. D. Singha, J. Kumari, B. Tiru, and U. Sarma, "Classification of soil types from GPR B scans using deep learning techniques," in *Proc. Int. Conf. Recent Trends Electron. Inf. Commun. Technol.*, 2021, pp. 840–844.

[14] N. N. Das and B. P. Mohanty, "Root zone soil moisture assessment using remote sensing and vadose zone modeling," *Vadose Zone J.*, vol. 5, no. 1, pp. 296–307, 2006.

[15] E. G. Njoku and D. Entekhabi, "Soil moisture remote sensing: State-of-the-science," *Proc. IEEE*, vol. 85, no. 8, pp. 1349–1375, 1996, doi: 10.1109/5.535743.

[16] M. Davis and G. Thompson, "Mapping soil moisture with electrical resistivity tomography," *Geophysical Prospects*, vol. 64, no. 5, pp. 1145–1160, 2023.

[17] R. Anderson and S. Watson, "Remote sensing of soil moisture using SAR," *Remote Sens. Rev.*, vol. 19, no. 2, pp. 285–305, 2023.

[18] L. Evans and P. Stone, "Synthetic aperture radar: Applications in agriculture," *Agricultural Remote Sens. Basics*, vol. 31, no. 1, pp. 95–115, 2023.

[19] U. Ozkaya, F. Melgani, M. B. Bejiga, L. Seyfi, and M. Donelli, "GPR B scan image analysis with deep learning methods," *Measurement*, vol. 165, 2020, Art. no. 107770.

[20] K. Tomiyasu, "Tutorial review of synthetic-aperture radar (SAR) with applications to imaging of the ocean surface," *in Proc. IEEE*, vol. 66, no. 5, pp. 563–583, 1978.

[21] M. S. Peek, A. J. Leffler, L. Hipps, S. Ivans, R. J. Ryel, and M. M. Caldwell, "Root turnover and relocation in the soil profile in response to seasonal soil water variation in a natural stand of utah juniper (juniperus osteosperma)," *Tree Physiol.*, vol. 26, no. 11, pp. 1469–1476, 2006.

[22] K. Wu et al., "A new drone-borne GPR for soil moisture mapping," *Remote Sens. Environ.*, vol. 235, 2019, Art. no. 111456.

[23] D. J. Daniels, "Ground penetrating radar," *IET*, vol. 1, 2004.

[24] I. Lunt, S. Hubbard, and Y. Rubin, "Soil moisture content estimation using ground-penetrating radar reflection data," *J. Hydrol.*, vol. 307, no. 1–4, pp. 254–269, 2005.

[25] N. Barkataki, A. J. Kalita, and U. Sarma, "Automatic material classification of targets from GPR data using artificial neural networks," in *Proc. IEEE Silchar Subsection Conf.*, 2022, pp. 1–5.

[26] N. Peplinski, F. Ulaby, and M. Dobson, "Dielectric properties of soils in the 0.3-1.3-GHz range," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 3, pp. 803–807, May 1995.

[27] D. Comite, A. Galli, S. E. Lauro, E. Mattei, and E. Pettinelli, "Analysis of GPR early-time signal features for the evaluation of soil permittivity through numerical and experimental surveys," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 1, pp. 178–187, Jan. 2016.

[28] E. Sejdić, I. Djurović, and L. Stanković, "Fractional fourier transform as a signal processing tool: An overview of recent developments," *Signal Process.*, vol. 91, no. 6, pp. 1351–1369, 2011.

[29] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE signal Process. Mag.*, vol. 8, no. 4, pp. 14–38, Oct. 1991.

[30] A. Benedetto and F. Benedetto, "Remote sensing of soil moisture content by GPR signal processing in the frequency domain," *IEEE Sensors J.*, vol. 11, no. 10, pp. 2432–2441, Oct. 2011.

[31] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola, "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data," *Remote Sens.*, vol. 7, no. 12, pp. 16398–16421, 2015.

[32] N. Barkataki, S. Mazumdar, B. Tiru, and U. Sarma, "Estimation of soil moisture from GPR data using artificial neural networks," in *2021 IEEE Int. Conf. Technol. Res. Innov. Betterment Soc.*, 2021, pp. 1–5.

[33] I. Giannakis, A. Giannopoulos, and C. Warren, "A machine learning-based fast-forward solver for ground penetrating radar with application to full-waveform inversion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4417–4426, Jul. 2019.

[34] J. Smith and J. Doe, "Utility of GPR in agriculture: A review," *J. Agricultural Phys.*, vol. 35, no. 2, pp. 123–139, 2023.

[35] C. Lee and Y. Kim, "Deep moisture profiling using GPR in large farms," *Int. J. Soil Sci.*, vol. 48, no. 4, pp. 401–420, 2023.

[36] R. Miller and H. Ford, "ERT for soil moisture estimation in large agricultural lands," *J. Environ. Eng. Geophys.*, vol. 25, no. 6, pp. 677–690, 2023.

[37] A. Brown and B. Green, "Assessment of soil moisture using time domain reflectometry," *J. Hydrol.*, vol. 120, no. 3, pp. 215–234, 2022.

[38] E. White and J. Black, "TDR and its application in precision agriculture," *Precis. Agriculture*, vol. 17, no. 1, pp. 58–76, 2022.

[39] Q. Liu et al., "Estimation of soil moisture using multi-source remote sensing and machine learning algorithms in farming land of northern China," *Remote Sens.*, vol. 15, no. 17, 2023, Art. no. 4214.

[40] D. D. Alexakis, F.-D. K. Mexis, A.-E. K. Vozinaki, I. N. Daliakopoulos, and I. K. Tsanis, "Soil moisture content estimation based on Sentinel-1 and auxiliary earth observation products. a hydrological approach," *Sensors*, vol. 17, no. 6, 2017, Art. no. 1455.

[41] V. Perez-Gracia, M. Solla, and S. Fontul, "Analysis of the GPR signal for moisture detection: Application to heritage buildings," *Int. J. Architectural Heritage*, vol. 18, no. 2, pp. 230–253, 2024.

[42] A. Annan, "Electromagnetic principles of ground penetrating radar," *Ground Penetrating Radar Theory Appl.*, pp. 1–40, 2005.

[43] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, no. 3, pp. 199–217, 2021.

[44] J. Shlens, "A tutorial on principal component analysis," 2014, *arXiv:1404.1100*.

[45] Y. Dai et al., "The common land model," *Bull. Amer. Meteorological Soc.*, vol. 84, no. 8, pp. 1013–1024, 2003.

[46] Y. Shin, B. P. Mohanty, and A. V. Ines, "Soil hydraulic properties in one-dimensional layered soil profile using layer-specific soil moisture assimilation scheme," *Water Resour. Res.*, vol. 48, no. 6, 2012.

[47] C. Warren, A. Giannopoulos, and I. Giannakis, "GprMax: Open source software to simulate electromagnetic wave propagation for ground penetrating radar," *Comput. Phys. Commun.*, vol. 209, pp. 163–170, 2016.

[48] D. M. Sullivan, *Electromagnetic Simulation Using the FDTD Method*. Hoboken, NJ, USA: Wiley, 2013.

[49] Natural Resour. Conservation Service, "A soil profile," Accessed: [Online]. Available: https://www.nrcs.usda.gov/resources/education-and-teaching-materials/a-soil-profile

[50] 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 Manage.*, vol. 245, 2021, Art. no. 106649.

[51] G. C. Topp, J. Davis, and A. P. Annan, "Electromagnetic determination of soil water content: Measurements in coaxial transmission lines," *Water Resour. Res.*, vol. 16, no. 3, pp. 574–582, 1980.

[52] OpenAI ChatGPT, "Synthetic GPR dataset visualization," [Image] Retrieved from conversation with OpenAI's ChatGPT, 2023, Accessed: Mar. 5, 2023.

[53] J. Hosken, "Ricker wavelets in their various guises," *First Break*, vol. 6, no. 1, 1988.

[54] J. D. Taylor, *Ultra-Wideband Radar Technology*. Boca Raton, FL, USA: CRC Press, 2000.

[55] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[56] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, pp. 1189–1232, 2001.

[57] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[58] B. Yegnanarayana, *Artificial Neural Networks*. PHI Learning Pvt. Ltd., 2009.

[59] X. L. Travassos, S. L. Avila, and N. Ida, "Artificial neural networks and machine learning techniques applied to ground penetrating radar: A review," *Appl. Comput. Informat.*, vol. 17, no. 2, pp. 296–308, 2020.

[60] A. Gelman, "Analysis of variance–why it is more important than ever," 2005.

[61] L. St et al., "Analysis of variance (ANOVA)," *Chemometrics Intell. Lab. Syst.*, vol. 6, no. 4, pp. 259–272, 1989.

[62] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results," *Int. J. Comput. Appl.*, vol. 44, no. 9, pp. 875–886,Author: Please check and confirm whether the authors affiliation in the first footnote are correct as set. 2022.

[63] M. Hashemi and H. Karimi, "Weighted machine learning," *Statistics, Optim. Inf. Comput.*, vol. 6, no. 4, pp. 497–525, 2018.

[64] S. Pathirana, S. Lambot, M. Krishnapillai, M. Cheema, C. Smeaton, and L. Galagedara, "Ground-penetrating radar and electromagnetic induction: Challenges and opportunities in agriculture," *Remote Sens.*, vol. 15, no. 11, 2023, Art. no. 2932.