Comprehensive GPR Signal Analysis via Descriptive Statistics and Machine Learning

Himan Namdari*, Majid Moradikia*, Douglas Todd Petkie* , Radwin Askari†, Seyed Zekavat*
* Worcester Polytechnic Institute (WPI), Worcester, MA, USA
Email: {hnamdari, mmoradikia,dtpetkie, rezaz}@wpi.edu
† Michigan Tech Research Institute (MTRI), Ann Arbor, Michigan, USA
Email: {raskari}@mtu.edu

***Abstract*—This paper presents a comprehensive analysis of how various soil characteristics impact the features of Ground Penetrating Radar (GPR) received signals. These characteristics include dielectric properties, thickness, number of layers, radar configuration, and surface roughness. The paper conducts an exhaustive analysis using gprMax, simulating diverse soil medium scenarios to demonstrate how these parameters influence the GPR-received signals. The proposed methodology extracts critical features from the received signal for soil characterization through descriptive statistical analysis. The paper then deploys Machine Learning (ML) techniques, specifically a Random Forest (RF) model and Gini Mean Decrease Impurity (MDI) as measures, to identify the most influential features in the dataset. This process extracts a concise set of features from the time domain, followed by an expansion using frequency domain features. The proposed approach not only effectively captures the critical information in the high-dimensional GPR data but also reduces its dimensionality, ensuring the preservation of essential information. Training ML and Deep Learning (DL) models using these significant features, rather than complex raw A-scan data, leads to more accurate soil moisture and subsurface analysis.**

***Index Terms*—GPR, signal processing, machine learning, feature extraction, soil characterization.**

I. Introduction

Ground-penetrating radar (GPR) is a non-invasive geophysical method used for subsurface imaging and analysis in various fields such as agriculture [1], civil engineering [2], and archaeology [3]. The performance of GPR in detecting objects and characterizing soil depends primarily on different subsurface parameters. These parameters include the soil's relative permittivity (ability to store electrical energy), electrical conductivity (ability to conduct an electric current), texture (relative proportions of sand, silt, and clay), soil moisture, and composition of the soil. Additionally, factors such as the soil's rough-surface profile [4], underground anomalies, operating frequency [5], and subsurface objects [6] influence the behavior of GPR waves as they traverse the soil. These factors affect the waves' speed, reflection, and refraction, modifying the power and shape of the received signals resulting in a more complex, high-dimensional, and diverse dataset [7].

Among these factors, relative permittivity is critical as it directly impacts other soil properties, thereby influencing the overall performance of the GPR signal. Thus, understanding and accounting for the interconnected nature of these variables is crucial [8]. Therefore, researchers have begun to leverage machine learning (ML) techniques, known for their powerful predictive and analytic capabilities, as a robust way to interpret complex interactions and dependencies within GPR data. However, such models typically perform well when trained on large amounts of low-dimensional data [9].

To convert high-dimensional GPR data into a format more suitable for ML models, it is necessary to first extract features from the received GPR signals[10]. Various techniques can be employed for this purpose, including Principal Component Analysis (PCA), Independent Component Analysis (ICA), Autoencoders, and t-distributed Stochastic Neighbor Embedding (t-SNE). These techniques involve mathematical procedures that transform the original high-dimensional data into a lower-dimensional space while preserving essential information. However, these methods have limitations, e.g., PCA and ICA might not effectively capture non-linear structures in soil-related GPR data and are sensitive to variable scaling [11].

In our study, we address these concerns with a new approach. Rather than reducing the dimensionality of GPR data directly and losing critical information, we employ descriptive statistics to extract key features from the received signal. This approach allows us to capture the non-linearity and inherent complexities of the soil-related GPR data. Afterward, we use the Random Forest (RF) model to select the most robust features less sensitive to variable scaling. The process begins by extracting a limited set of features from the time domain. This set is then expanded using features from the frequency domain.

Our proposed methodology effectively captures the critical information of the GPR data and reduces its dimensionality. Thus, using these critical features to train ML and deep learning (DL) models instead of the complex raw A-scan data results in more accurate subsurface analysis. In summary, our work contributes to the field by analyzing GPR data characteristics, extracting key features using descriptive statistics and a Random Forest model, and ranking these features based on their importance. This approach enhances the performance of subsequent machine learning and deep learning models in analyzing subsurface structures.

This work was supported by the United States Department of Agriculture (Grant Number: USDA NR223A750013G032). We gratefully acknowledge their financial support for this research.

II. Preliminaries and Related Work

Recently, researchers have employed various methods in both time and frequency domains to analyze GPR received signals, aiming to estimate or categorize soil characteristics. While the time domain provides direct data interpretation useful for detecting layers and estimating depth, the frequency domain offers more details about subsurface material properties like permittivity texture and porosity [12, 13, 14]. Thus, combining both domains could enhance data interpretation, improve feature accuracy, and facilitate a more comprehensive analysis.

Soil characteristics influence the propagation speed of the EM signal, thereby affecting the GPR signal's velocity, peak amplitude, and time-of-arrival (ToA) [15]. The velocity, v_{soil}, is calculated using the formula $v_{soil} = c/\sqrt{\epsilon\mu}$, where c is the light speed, ϵ stands for soil permittivity, and μ denote the relative permeability [15]. Inhomogeneous soil with known permittivity, ϵ_{soil}, the signal's ToA of the GPR signal is given by $(ToA) = 2d/v_{soil}$, with d representing soil's depth [15]. The GPR reflection coefficient (ratio of reflected incident signal amplitude at a soil boundary) primarily depends on soil permittivity. By investigating this coefficient, we can distinguish soil boundaries based on their permittivities [16]. The reflection coefficient (R) is defined as follows:

$$R = \frac{\sqrt{\epsilon_1} - \sqrt{\epsilon_2}}{\sqrt{\epsilon_1} + \sqrt{\epsilon_2}} \tag{1}$$

Where ϵ_1 and ϵ_2 are the permittivities of the two different materials (e.g., air and soil).

Overall, permittivity directly impacts the previously mentioned properties and considerably affects the form and quality of the received signal. Such changes can consequently modify the soil properties. For example, soil's moisture content (SMC) is intrinsically linked to dielectric permittivity and commonly calculated through petrophysical correlations, as shown by Topp's equation [17], as follows:

$$\theta_{soil} = -5.3\times10^{-2} + 2.92\times10^{-2}\epsilon_{soil} - 5.5\times10^{-4}\epsilon_{soil}^2 + 4.3\times10^{-6}\epsilon_{soil}^3 \tag{2}$$

Here, θ_{soil} denotes soil moisture (Volumetric Water Content), and ϵ_{soil} represents the soil permittivity value.

Therefore, a precise permittivity estimation can enhance the accuracy of soil moisture prediction, especially when using ML models. GPR analysis utilizing ML typically uses complex A-scan (1D) data or B-scan (2D) as input to train ML and DL models [18]. This method can be computationally demanding due to the size and complexity of input data, particularly with GPR data, which often involves large datasets. A possible solution to this challenge is to use statistical feature extraction. This method simplifies the input data and improves efficiency, robustness, and interpretability, making it less computationally demanding. For instance, Smitha et al. [19] proposed a novel method using supervised ML, where they extracted features like mean, variance, kurtosis, skewness, and entropy to train Support Vector Machine (SVM) and Neural Network (NN) models for improving the accuracy of object detection. Similarly, authors in [20] extracted statistical features to rapidly diagnose moisture damage in asphalt pavement. The model demonstrated high accuracy, significantly improving asphalt quality evaluation.

ML model requires a lot of data for training. However, gathering GPR data on-site for these methods is costly, time-consuming, and logistically difficult due to field survey requirements. In practice, various environmental factors will influence the collected data, leading to uncertainties and increased noise, making it less suitable for model training. To circumvent these limitations, simulators such as gprMax[21] can be leveraged to generate synthetic GPR data. This synthetic data can emulate real-world scenarios under controlled conditions. Recent studies have explored gprMax and its capability of generating synthetic GPR data. For instance, authors in [22] propose a three-step method for locating and identifying underground objects using gprMax synthetic data. The proposed method boosts Overall performance from 80.4% to 90.3% using K-Nearest Neighbors (KNN) with K=5. However, the detection performance is slightly lower than the Histograms of Oriented Gradients (HOG) method. In another paper, [23], the author used gprMax-generated synthetic data to estimate moisture in asphalt concrete. The results validated the presented model and, therefore, demonstrated the ability of GPR to monitor moisture variation in AC pavements. These papers only focused on a specific characteristic of soil and task and failed to fully assess the versatility of gprMax in managing various models. They also used the data for their tasks rather than conducting an in-depth analysis of all soil properties. Therefore, a comprehensive evaluation is crucial to understand the full potential of gprMax in soil analysis.

Our proposed model addresses traditional methods' limitations with a simplified and effective feature extraction technique. We simulated and individually labeled diverse soil scenarios using the gprMax simulator, considering multiple factors impacting GPR signals. This approach facilitates comprehensive feature analysis and evaluates the significance of these features in univariate or multivariate simulated models, identifying the most essential features that will be leveraged for ML training proposes.

III. Methodology and Experimental Setup

We used the gprMax software [21] to simulate diverse soil models and their corresponding GPR EM waves via a finite-difference time-domain (FDTD) method [24]. Inputs such as model size, discretization, antenna parameters, soil texture, and radar height facilitated diverse soil model creation with varying layers, permittivity, roughness, and depths Fig. 1. The gprMax generated time-domain A-scan signals, and we customized the gprMax input code to save the labels (e.g., permittivity, depth, and the number of layers) separately Fig. 2. We utilized Python and HDF5 to analyze the Ez element (Ez represents the component of the electric field in the z-direction. The positive or negative sign of Ez indicates the direction along the z-axis) of the electric field vector E, extracting the time domain signal. This signal merged with the labeled dataset and experienced feature extraction using Python. To identify critical features, we employed a Random Forest algorithm. Fig. 3 provides an example of the A-scan data.

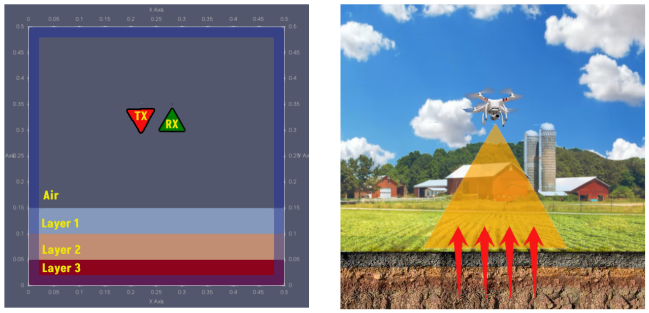

Fig. 1. Figure shows the real (right) and simulated (left) environment, including the GPR UAV in real and simulated radar over farmland and simulated soil with three soil layers with different permittivities.

Our simulation featured a domain size of 1m × 1m × 0.002m, using a 1.5 GHz central frequency Ricker wave. This waveform emulates real GPR system pulses, like those from Geophysical Survey Systems, Inc (GSSI), allowing for high-resolution time-domain data. We recorded the signal for five ns, the transmitter, and receiver, placed 0.08m apart utilizing fix offset setup and 1.0 (m) above the surface, emulated an Air-coupled GPR radar setup Fig. 1. Soil moisture fluctuation was simulated by changing soil layer permittivity from 5 to 60 in increments of 5, representing conditions from dry (ϵ = 5) to wet (ϵ = 60) soil. Following gprMax guidelines, we set the model's time window to 5ns, yielding an 1161 x 1 vector representing signal amplitude at corresponding sample points Fig. 3. We further modeled 1-3 layers, adjusting layer depth between 0.1 and 0.3m Fig. 1. Additional parameters like surface roughness, signal central frequency, and radar height were also explored. Due to space constraints, we detail permittivity, layer depth, and layer number variations, while other experimental observations are briefly discussed in the results section.

A. Feature Extraction and Feature Analysis

Feature extraction is crucial for identifying subsurface targets and interpreting soil characteristics in GPR data. The descriptive analysis involves summarizing data using a few key statistical metrics. The primary goal is to describe the main features of the collected data in a quantitative manner. In our study, we used descriptive analysis on A-scan GPR data to extract key signal amplitude-related features such as min/max, mean, standard deviation, quartiles, mean peaks, kurtosis, and skewness. We also applied a Fast-Fourier Transform (FFT) to the raw signal to extract frequency domain features such as FFT max (peak frequency), FFT mean (average power), and FFT std (power variability). These time and frequency-domain features offer crucial insights into the nature of the subsurface captured by the GPR signal. This approach improves the interpretability of GPR data analysis; unlike PCA or ICA, which aim at reducing the dimensionality of data, descriptive analysis keeps the original variables intact. Descriptive analysis can handle non-linear relationships between variables, whereas PCA and ICA are linear techniques.

B. Feature Importance

Feature importance serves as a quantitative measure, signifying the relative value of each feature in machine learning models, particularly in relation to predicting the target variable [25]. In our study, we employed the Random Forest model coupled with Gini Mean Decrease Impurity (MDI) to evaluate the significance of the extracted features. The result of the feature importance analysis can be found in Fig. 7

$$Gini(node) = 1 - \sum_{i=1}^{k} (p_i)^2 \tag{3}$$

Where p_i is the proportion of samples of class i in the node, and the feature importance is as follows:

$$(Gini) = \sum_{i=1}^{n} \frac{(Gini(parent) - Gini(child))}{Total\ number\ of\ trees} \tag{4}$$

The more significant the impurity decrease, the more influential the feature is, as seen in Fig. 7.

C. GPR Features Correlation Matrix

A correlation matrix offers a comprehensive view of the pairwise relationships among statistical features derived from A-scan GPR signals, revealing feature interdependencies. The matrix's coefficients represent both the strength and direction of these relationships: a high positive coefficient suggests a strong direct association, while a high negative one indicates a strong inverse relationship. A coefficient near zero, however, denotes a weak or non-existent association. Our result of features correlation analysis can be found in Fig. 8.

IV. Experimental Results

Our results for different experiments are as follows: In the permittivity benchmark with one layer, we observed that increasing the permittivity resulted in higher signal attenuation, delay, and reduced signal velocity Fig. 2.

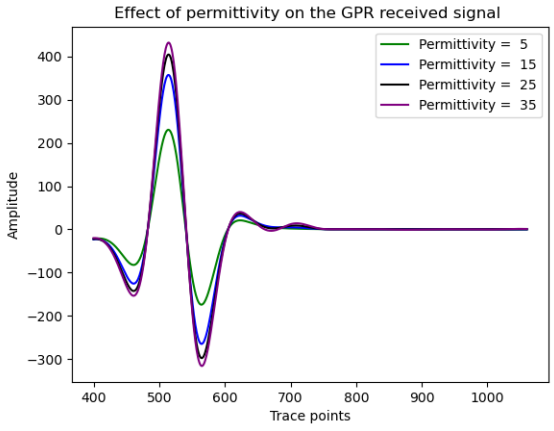

Fig. 2. The impact of permittivity on the received signal with 1061 samples equal to five ns of recorded data (To enhance visualization, we isolated the A-scan by segmenting sample points [0-400], the direct wave between the transmitter and receiver (zero time correction or ZTC).

Moreover, in the multi-layer benchmark, the results showed that the depth of the soil could impact the Time of Arrival

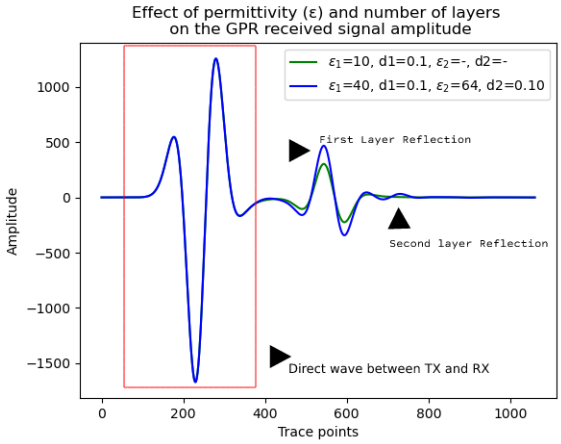

Fig. 3. Comparison of signal behavior in one and two-layer models. In the two-layer model, the signal undergoes two reflections at the layer interfaces. This leads to wavefront distortion and results in more complex signal characteristics compared to the one-layer model.

(ToA) and the number of peaks in the signal. In the one-layer model, we observed two peaks, while in the two-layer model, there were three peaks, including the direct wave between TX and RX. The index of these peaks in the E vector can be converted to estimate the signal delay and depth and the number of existing layers Fig. 3.

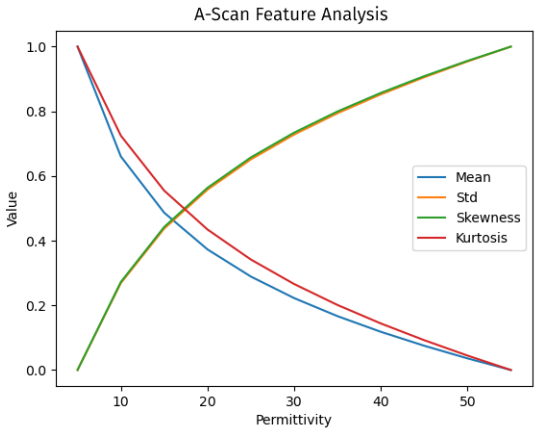

Fig. 4. Permittivity Variation impact on mean, skewness, std, and kurtosis: Increasing permittivity corresponds to a gradual increase in skewness, std, and decrease in mean and kurtosis.

In the surface roughness experiment, we observed that uneven surfaces could cause scattering of the GPR signal, leading to a weaker return signal and potentially obscuring subsurface targets. In the frequency benchmark, we realized that lower frequencies, such as 0.5 GHz, can achieve greater penetration depth. However, it becomes more challenging to identify smaller targets or thin layers. On the other hand, higher frequencies like 1.5 and 2.0 GHz offer improved resolution, enabling the detection of smaller targets and finer subsurface. Increasing antenna separation will create higher delay and enhances GPR signal depth while decreasing it improves resolution for shallow targets. A higher UAV height expands ground coverage but reduces sensitivity and resolution, making detecting small or shallow targets difficult. Lowering UAV

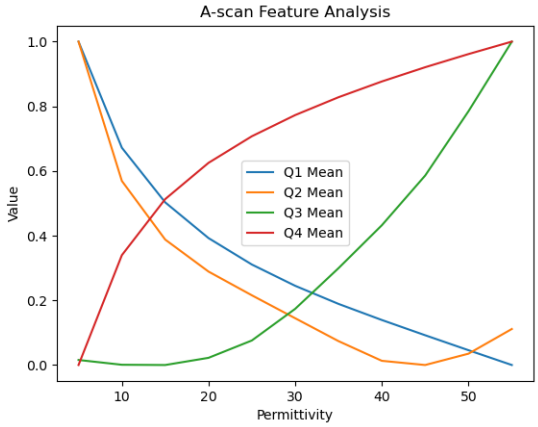

Fig. 5. Permittivity Variation impact on quartiles statistical features: Increasing permittivity corresponds to a gradual increase in Q1, Q2, and decrease in Q3 and Q4 mean.

height increases resolution and sensitivity but limits coverage due to longer flight times.

Considering these observations and the diverse impact of each feature, we proposed a descriptive-based feature extraction method. This approach can summarize and extract these impacts and, by reducing the input size, will provide clearer and less noisy training data for the ML model. In our expanded simulation, we adjusted permittivity according to changes in soil components, used the A-scan data, and extracted relevant features from these simulations. Our primary aim was to identify any noticeable and informative patterns among these features. The results indicate that as permittivity increases, features such as Q3, Q4 mean, std, skewness, peak mean, and FFT extracted features increase due to a positive correlation, while features like kurtosis, mean, and Q1,2 decrease due to a negative correlation Fig. 4, 5, 6.

Our extensive simulation focused on identifying the most robust features and examining their interrelations using a feature importance analysis. The results indicated high significance for many features, with Q4, skewness, kurtosis, and FFT

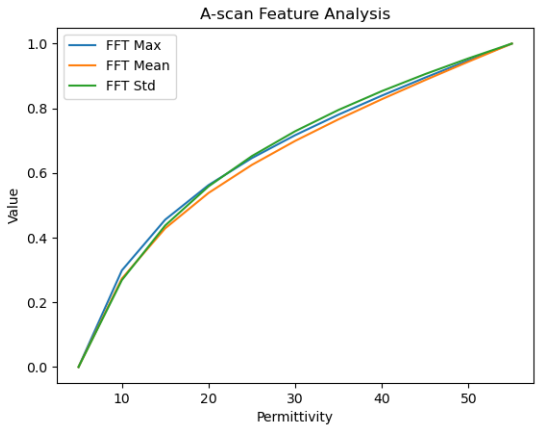

Fig. 6. Permittivity Variation impact on extracted FFT statistical features: Increasing Permittivity Corresponds to Gradual Increase in FFT's Mean, Maximum, and Standard Deviation.

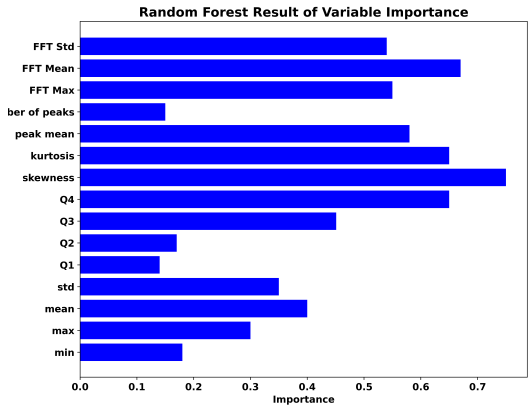

Fig. 7. Feature importance analysis of the statistically extracted features utilizing the random forest model.

mean standing out as most crucial, except for min, Q1, Q2, and the number of peaks Fig. 7. Furthermore, the correlation matrix highlighted a strong association between std, skewness, Q4, and FFT features, validating our findings from the feature importance analysis Fig. 8. The observed patterns in these features highlight their effectiveness in enhancing the understanding of signal behavior in relation to its diverse characteristics.

V. Conclusion

This research paper presented a comprehensive and thorough analysis of various soil characteristics and their impact on the features of GPR-received signals. The experimental investigation using gprMax allowed us to simulate different soil scenarios and observe how these variables affect GPR received signals. This rigorous exploration highlighted the significant impact of factors such as dielectric properties, layer thickness, and surface roughness on GPR signal characteristics. Using descriptive statistical analysis, we managed to convert high-dimensional, complex GPR data into a more tractable, lower-dimensional form. ML techniques, specifically Random Forest, were used to assess the importance of these features, enhancing our understanding of the GPR signals and paving the way for more accurate soil subsurface analysis. Simulation results show that permittivity, layering, surface roughness, and frequency of the soil considerably impact the GPR signals. As we varied permittivity, we noticed noticeable patterns among our extracted features, further validating our descriptive-based feature extraction method. These highly correlated features will be used in the training process of ML and DL models and improve the accuracy and lowers the dimentionality of the input data.

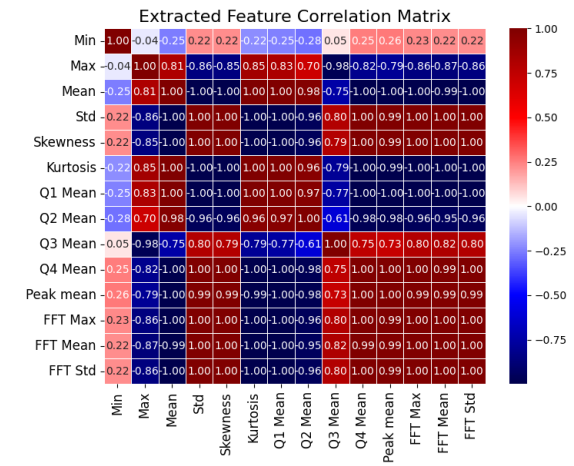

Fig. 8. Extracted feature correlation analysis. A high correlation between std and skewness with Q3, Q4 mean, and FFT extracted features.

References

[1] Alessandro Fedeli et al. “COST Action TU1208” Civil Engineering Applications of Ground Penetrating Radar:” a short overview of the main scientific activities and results.” In: *Geophysical Research Abstracts*. Vol. 21. 2019.

[2] Petri Linna, Antti Halla, and Nathaniel Narra. “Ground-Penetrating Radar-Mounted Drones in Agriculture”. In: *New Developments and Environmental Applications of Drones: Proceedings of FinDrones 2020*. Springer. 2022, pp. 139–156.

[3] Lara Pajewski, Mercedes Solla, and Melda Küçükdemirci. “Ground-Penetrating Radar for Archaeology and Cultural-Heritage Diagnostics-Activities Carried Out in COST Action TU1208”. In: *Nondestructive Techniques for the Assessment and Preservation of Historic Structures*. CRC Press, 2017, pp. 215–225.

[4] Magda El-Shenawee and Carey M Rappaport. “Quantifying the effects of different rough surface statistics for mine detection using the FDTD technique”. In: *Detection and Remediation Technologies for Mines and Minelike Targets V*. Vol. 4038. SPIE. 2000, pp. 966–975.

[5] Daniela De Benedetto et al. “Impact of data processing and antenna frequency on spatial structure modelling of GPR data”. In: *Sensors* 15.7 (2015), pp. 16430–16447.

[6] Levent Gurel and Ugur Oguz. “Simulations of ground-penetrating radars over lossy and heterogeneous grounds”. In: *IEEE Transactions on geoscience and remote sensing* 39.6 (2001), pp. 1190–1197.

[7] Timothy W Miller, JMH Hendrickx, and B Borchers. “Radar detection of buried landmines in field soils”. In: *Vadose Zone Journal* 3.4 (2004), pp. 1116–1127.

[8] Jan MH Hendrickx, Bhabani S Das, and Brian Borchers. “Modeling distributions of water and dielectric constants around land mines in homogeneous soils”. In: *Detection and Remediation Technologies for Mines and Minelike Targets IV*. Vol. 3710. SPIE. 1999, pp. 728–738.

[9] Xisto L Travassos, Sérgio L Avila, and Nathan Ida. “Artificial neural networks and machine learning techniques applied to ground penetrating radar: A review”. In: *Applied Computing and Informatics* 17.2 (2020), pp. 296–308.

[10] Brian Karlsen et al. “Comparison of PCA and ICA based clutter reduction in GPR systems for anti-personal landmine detection”. In: *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing (Cat. No. 01TH8563)*. IEEE. 2001, pp. 146–149.

[11] Pawel Kaczmarek and Jerzy Pietrasinski. “Principal component analysis in interpretation of A-Scan measurements in GPR system”. In: *2014 15th International Radar Symposium (IRS)*. IEEE. 2014, pp. 1–5.

[12] Peter King-Wah Lau et al. “Characterizing pipe leakage with a combination of GPR wave velocity algorithms”. In: *Tunnelling and Underground Space Technology* 109 (2021), p. 103740.

[13] Kenneth OM Mapoka et al. “Using gprMax to Model Ground-Penetrating Radar (GPR) to Locate Corn Seed as an Attempt to Measure Planting Depth”. In: *Transactions of the ASABE* 62.3 (2019), pp. 673–686.

[14] Iurianne MM Conti et al. “Porosity estimation and geometric characterization of fractured and karstified carbonate rocks using GPR data in the Salitre Formation, Brazil”. In: *Pure and Applied Geophysics* 176 (2019), pp. 1673–1689.

[15] Harry M Jol. *Ground penetrating radar theory and applications*. elsevier, 2008.

[16] Sébastien Lambot et al. “Analysis of air-launched ground-penetrating radar techniques to measure the soil surface water content”. In: *Water resources research* 42.11 (2006).

[17] G Clarke Topp, JL Davis, and Aa P Annan. “Electromagnetic determination of soil water content: Measurements in coaxial transmission lines”. In: *Water resources research* 16.3 (1980), pp. 574–582.

[18] Zi Xian Leong and Tieyuan Zhu. “Direct velocity inversion of ground penetrating radar data using GPRNet”. In: *Journal of Geophysical Research: Solid Earth* 126.6 (2021), e2020JB021047.

[19] N Smitha and Vipula Singh. “Target detection using supervised machine learning algorithms for GPR data”. In: *Sensing and Imaging* 21.1 (2020), p. 11.

[20] Jun Zhang et al. “In-situ recognition of moisture damage in bridge deck asphalt pavement with time-frequency features of GPR signal”. In: *Construction and Building Materials* 244 (2020), p. 118295.

[21] Craig Warren, Antonios Giannopoulos, and Iraklis Giannakis. “gprMax: Open source software to simulate electromagnetic wave propagation for Ground Penetrating Radar”. In: *Computer Physics Communications* 209 (2016), pp. 163–170.

[22] İbrahim Meşecan, Betim Cico, and İhsan Ömür Bucak. “Feature vector for underground object detection using B-scan images from GprMax”. In: *Microprocessors and Microsystems* 76 (2020), p. 103116.

[23] Qingqing Cao and Imad L Al-Qadi. “Effect of moisture content on calculated dielectric properties of asphalt concrete pavements from ground-penetrating radar measurements”. In: *Remote Sensing* 14.1 (2021), p. 34.

[24] Kurt L Shlager and John B Schneider. “A selective survey of the finite-difference time-domain literature”. In: *IEEE Antennas and Propagation Magazine* 37.4 (1995), pp. 39–57.

[25] Bjoern H Menze et al. “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data”. In: *BMC bioinformatics* 10 (2009), pp. 1–16.